# ASSESSING PAIN THROUGH BEHAVIOURAL OBSERVATION

by

Kelly Anne Wade

A THESIS SUBMITTED TO THE UNIVERSITY OF BIRMINGHAM FOR THE DEGREE OF DOCTOR OF CLINICAL PSYCHOLOGY

Department of Clinical Psychology

School of Psychology

The University of Birmingham

June 2019

Dedicated to my Dad, whose care for others set me on this path,
and to my Mum, whose support kept me on it.

(And to my sister Alex for keeping me company along the way).

## Acknowledgements

This project would not have come together without the considerable help and support that I have received from a number of people.

A very special thank you to Chloe Brown who helped me locate the data, and spent valuable time coding videos and screening papers. Without you a project of this size would not have been possible. To Dr Chris Jones and his support with the meta-analysis including his very useful and clever R script, thank you. To Dr Caroline Richards, who has been much more than a supervisor to me, who has, and continues to be, a role model and a constant source of support and inspiration. Thank you all.

A huge thank you also to all of the researchers and students of the Cerebra Centre who have gone before me and the contributions they have made to the projects that provided the data for this thesis. Thank you also to the children and families who participated in those projects. In particular thank you to Prof Chris Oliver, Dr Lucy Wilde, and Dr Kate Eden for the support I received during my MRes, it was that experience that truly inspired this.

My thanks also to Dr Isabel Clare, Dr Adam Wagner and Prof Tony Holland of Cambridge Intellectual and Developmental Disabilities Research Group, who helped me develop the skills and confidence that have enabled me to take on a thesis project that feels meaningful and worthwhile.

To all those outside of the university campus: to Jo and Matt, whose company and distraction has never before been more valued than it has in the last 12 months! To Shona, Jade, Rosie, and Chloe, who are the world’s best cheering squad, no matter what the stage. To my family, whose support has never wavered, there are not enough words to express my feelings. Nothing brings to light the importance and influence of families quite like clinical training, and I have been blessed with the best in the world. And finally, to Dave, who has more faith in me than I have in myself. Thank you. Here’s to the next chapter…

# Thesis Overview

**Background:** The health inequalities faced by people with intellectual disabilities (ID) are well documented, affecting both duration and quality of life. Painful health conditions can be difficult to recognise as many people with ID struggle to self-report their pain. Therefore it is important that accurate observational tools are available to support recognition and assessment of pain in people with ID.

**Aim:** This thesis seeks to assess the use of currently available observational assessments of pain through meta-analytic methods, and then evaluates a more specialist observational tool designed to detect gastric pain.

**Meta-analysis:** A comprehensive review of the literature found 62 distinct observational measures used in published research. The five most commonly used measures were assessed through a series of meta-analyses, synthesising correlations between observational and self-report measures of pain. Moderate correlations were found for all observational measures compared to self-report, though unacceptable levels of heterogeneity were also identified. Recommendations are made for use of the Face, Legs, Activity, Crying, Consolability scale.

**Empirical paper:** The Gastric Distress Questionnaire (GDQ) is a parent report measure designed to screen for Gastro-oesophageal Reflux Disease (GORD), a painful health condition which is common in people with ID. Significant differences in GDQ scores were found between children with and without recent GORD. A clinical cut off is recommended for the use of the GDQ to screen for reflux.  Behavioural observation by a naïve observer was not found to associate to GDQ scores provided by a parent, emphasising the importance of caregiver report in identifying GORD.

## Contents Volume One

## List of Figures

## List of Tables

## Contents Volume Two

# CHAPTER ONE, LITERATURE REVIEW:

# OBSERVATIONAL ASSESSMENTS OF PAIN IN CHILDREN AND ASSOCIATION WITH SELF-REPORT: A META-ANALYSIS

## 1.1. Abstract

**Background:** Due to the subjective nature of pain, self-report is widely accepted as the gold standard assessment, even in young children. However, there are many people who cannot give self-report ratings for pain. When self-report is not feasible, observational measures are often used to provide a proxy rating of pain. This review and meta-analysis seeks to explore the use of observational measures of pain in children in the research literature, and to evaluate their validity through correlation to self-report.

**Method:** Five databases were searched, yielding an initial return of 18335 papers. Once the papers had been screened, 526 studies were identified that reported pain scores obtained from observational measures in children aged 1-18 years. A series of meta-analyses were conducted to synthesise published findings on the associations between any of the five most common measures and self-report of pain.

**Results:** Sixty two different measures were identified, but 346 of the reviewed papers reported using either the Children's Hospital East Ontario Pain Scale (CHEOPS; McGrath et al., 1985), the Visual Analogue Scale (VAS; Price, McGrath, Rafii & Buckingham, 1983), the Face, Legs, Activity, Crying and Consolability (FLACC; Voepel-Lewis, Shayevitz & Malviya, 1997), the Observer Pain Scale (OPS; Hannallah et al., 1987), and the Wong-Baker Faces Scale (WBF; Wong & Baker, 1988). All five of the measures were found to have a moderate correlation to self-report.

**Conclusion:** There is an excess of observational measures being used in the research, without sufficient evidence to support their use. The FLACC is

recommended as a useful structured assessment of pain with moderate correlation to self-report.

## 1.2. Introduction

Acute pain is widely considered to be the human body's "warning signal", typically triggered by injury or disease (Loeser & Melzack, 1999). Pain is a universal human experience that all children will experience many times as they grow and learn about the world. In their early years, most children undergo many painful routine medical procedures such as vaccination, and a substantial minority will also undergo other painful procedures such as surgery. There is a drive in the medical literature to develop procedures that are less painful, and to improve the efficacy of pain management (Stinson et al., 2008). However, the improvement of procedures, and the success of routine pain management is dependent upon recognition and accurate assessment of pain (Verghese & Hannallah, 2010).

Accurate assessment of pain is a complex process. Current models of pain recognise pain as subjective, being influenced by psychological components of cognition and emotion, such as anxiety, expectations and context, and early experience of injury (Merksey, 1991; Craig, 2009). Given the subjective nature of pain, self-report is the preferred method of pain assessment in adults and for children as young as three years old (Royal College of Nursing, 2009). However, not all children are able to engage with self-report measures. Self-report of pain requires an understanding of the concept of pain, an ability to identify and label internal states, and a comprehension of magnitude and serial order (Fanurik, Koh, Harrison, Conrad & Tomerun, 1998). Young children and children with cognitive impairment, communication impairments, or those in temporary states of distress or confusion may be unable to provide accurate self-reports of pain (Voepel-Lewis, Malviya & Tait, 2005). The consequences of failing to recognise pain can be severe. Experiences of pain have been connected, even in the short term, to reduced affect and an increased need to aggress (Riva, Wirth & Williams, 2011). In people with cognitive impairments, untreated pain is associated with behaviours that challenge, in particular self-injurious behaviour (Carr & Owen-DeSchryver, 2007), which can have a significant negative impact on quality of life (Beadle-Brown, Murphy & DiTerlizzi, 2009). It has also been proposed that whilst self-report is an important tool to capture a subjective experience, it is merely one aspect of a more complex behavioural response to pain (Anand & Craig, 1996). Thus, despite self-report being identified as the gold-standard of pain assessment, it is neither adequate nor appropriate in all cases.

In cases where self-report is not feasible or appropriate, pain assessments are frequently carried out based upon the judgements of others. Von Baeyer & Spagrud (2007) describe four primary groups of observational pain tools applied in research and clinical practice. ‘Behavioural checklists’ and ‘behaviour rating scales’ both require observers to identify and/or score a series of pain related behaviours that may be observed in the child in question, to produce a total score indicative of pain intensity. ‘Global behaviourally anchored rating scales’ also make use of behavioural indicators, but descriptions of behaviours are given only as examples to guide the rating, observation of specific behaviours is not rated or required to justify a rating. Finally, ‘global rating scales’ require the observer to provide a rating of pain based on their own judgement without any reference to specific behaviours. Global rating scales appear to rely upon the implicit human ability to recognise pain in others and, despite being widely used, they have been criticised for being oversimplified and failing to capture the complexity of the pain response (Williams, Davies & Chaudry, 2000). Given the range of observational tools available, and the application of such tools as a proxy rating when a person is unable to self-report (e.g., young children, children with cognitive impairment), it is essential that the validity of these observational measures is evaluated.

Despite their widespread use, there is a lack of synthesised evidence regarding the psychometric properties of many common observational measures used to assess pain in children. A systematic review of observational measures of pain conducted by Von Baeyer and Spagrud (2007) identified only seven measures that were deemed to have sufficient evidence of validity and reliability to warrant recommending for use in clinical trials. However, a more recent systematic review by Andersen, Langius-Eklöf, Nakstad, Bernkley and Jylli (2017) identified a total of 65 observational assessments of pain cited across twelve published reviews of the research literature. Thus, there is inconsistency between the practice that is supported by evidence and the measures that are reportedly being used in current research. These data require reviewing and statistical synthesis to evaluate and improve current practice. Meta-analytic methods have not yet been applied to this literature. A synthesis of the available data regarding putative associations between observational methods and self-report will enhance the understanding of the validity of tools in current use.

Therefore, the present study seeks to extend the literature by providing quantitative evaluation of the use of observational pain assessments for children in published research, addressing two key aims:

1. To identify the most common observational assessments used in published studies that seek to quantify pain in children
2. To evaluate the validity of the most common observational assessments, by evaluating the association between pain scores obtained by observation and those obtained by the gold standard method, self-report.

## 1.3. Method

### 1.3.1. Literature Search

Ovid PsycInfo, Medline, Embase, Web of Science and CINAHL databases were searched for articles relating to the assessment of pain in children on 15th January 2018, using the search terms outlined in Table 1. As the aim of the initial search was to identify all research papers which made use of observational measures of pain, the search terms were broad. The search was restricted to the title and abstract and the terms relating to “pain” and “scale” were specified to be no more the three words apart from each other. Terms referring to “observation” or “behaviour” were not used for the initial search because a preliminary review of the literature identified that many papers used terms such as “pain scales” in the title or abstract and only clarified the use of observational scales in the body of the paper. Hand searches of returned reviews and meta-analyses were also conducted.

*Table 1*: Search terms used to identify papers relevant to pain assessment in children, applied to Ovid PsycInfo, Medline, Embase, Web of Science and CINAHL databases.

| Construct | Search terms |
|---|---|
| Pain | Pain*; Discomfort |
| Scale | Measure*; Scale*; Test*; Rating*; Assess*; Checklist* |
| Child | Child*; Adolescen*; Youth*; Infan*; Teen*; Juvenile*; Paediatric*; Pediatric* |

### 1.3.2. Inclusion Criteria

After duplicates were removed, titles and abstracts were reviewed to identify papers that met exclusion criteria. To maintain the breadth of the search and avoid premature exclusion of potentially useful papers, any papers that did not explicitly meet one or more exclusion criteria were retained to be screened at full paper stage. This approach was taken as a strategy to address poor reporting of samples in paper abstracts, for example, in many cases the age range of the sample was not reported in the abstract and therefore it was impossible to determine if studies met the inclusion criteria for age until the paper had reached full screen. In the initial paper screening the methodology and results sections were reviewed to identify those papers which met the inclusion criteria outlined in Table 2. For studies that met the inclusion criteria, the observational tools used, sample size, and sample age range were recorded.

A second screening stage was undertaken once all reported observational measures of pain had been identified. Papers reporting any of the ten most commonly used measures were screened for their inclusion of self-report measures of pain, and correlation between self-report and observational pain scores. The number of measures entered into this screen was chosen post-hoc as together they accounted for the vast majority of the literature (80.98%). The number of studies reporting correlations between self-report and observational pain scores was used to judge the feasibility of meta-analysis for each measure and determine which measures to include in the final meta-analysis.

*Table 2*: Inclusion and exclusion criteria applied to literature search at title and abstract screening, screening of methodology and results, and screening of full text for selection of primary studies for meta-analysis

| Inclusion Criteria | Justification | Exclusion criteria |
|---|---|---|
| Initial literature search – (i) title and abstract screening; (ii) initial screening of methodology and results | | |
| Assessment of pain | | |
| Methodology refers to a rating of pain intensity that is obtained by an observer. If a measure is cited which may be used either as self-report or observer report, the wording describing who makes the rating must be clear and unambiguous. | This is to ensure that studies using only self-report of pain are not included in error. Specialist measures were excluded so as to restrict the search to the measures with the broadest application and relevance. | Pain only assessed by self-report<br>• Only pain measures reported are described as “subjective” or being obtained through self-report<br>• Only pain measures reported are typically self-report measures e.g. Wong Baker Faces scale, with no explicit clarification given that score was determined through observation rather than self-report<br>Specialist measure<br>• Use of a measure that is designed to assess the pain associated with a specific health condition<br>• Use of a measure that is designed to assess pain in one specific body part or region |
| It must be clearly established that there is reason to believe the child is experiencing acute pain; the cause of pain should be established and described in the methodology | This is to limit the chances that the assessment of pain is being confounded by other emotional states such as anxiety or distress. | Pain not an outcome measure<br>• No attempt to quantify intensity of pain<br>• Assessment of associated construct e.g. distress or anxiety rather than pain<br>• No description given of painful procedure or incident to justify assessment of acute pain |
| Observational assessment used as an outcome measure | | |
| The observational measure of pain must be referenced in the results section of the study | This is to maintain the aim of the review in reviewing tools used as research measures, by excluding studies that cite the use of measures only as part of routine clinical practice without conducting any analysis of the ratings | Pain not an outcome measure<br>• No reference to pain scores or pain measures included in the results section |

| Inclusion Criteria | Justification | Exclusion criteria |
|---|---|---|
| | obtained. | |
| Pain in children<br>The pain is being observed in healthy human subjects who are between one and eighteen years of age | To maintain the focus of the review which is pain in children, and not potentially confound results with pain measures intended for infants or adults. | No participants <18 years<br>• Sample age range is all above 18 years<br>• Participants are explicitly described as “adult”<br>No participants >12 months<br>• Sample age range is below 12 months<br>• Participants are explicitly described as “newborn” “neonates” or “infants”<br>Unable to extract data for 1-18 year olds<br>• Sample includes children and adults, or children and infants, but is not stratified by age in results section<br>• Either upper or lower age limit is not given making it impossible to ascertain age range of sample<br>Chronic pain/health problem<br>• Sample selected for chronic, functional, or recurring pain<br>• Sample selected for chronic health problem associated with pain e.g. arthritis, fibromyalgia, cancer<br>Non-human subjects<br>• Animal studies |
| Research literature<br>The study is a quantitative or mixed methodology empirical study published in English in a peer reviewed journal | To maintain the focus of the review on use of measures in the research literature. | Not in English language<br>• Full text is not available in English<br>Not a quantitative empirical paper<br>• Case study<br>• Qualitative study<br>• Commentary<br>• Dissertation<br>• Protocol for proposed study<br>• Professional guidelines<br>• Book |

| Inclusion Criteria | Justification | Exclusion criteria |
|---|---|---|
| | | Review/meta-analysis<br>• Cochrane review<br>• Narrative literature review<br>• Other synthesis of literature |
| Meta-analysis – full text screening | | |
| Chosen observational measure<br>The study includes one of the chosen frequently used observational measures, as identified by the literature search and second screen | To ensure the meta-analysis is conducted on the observational measures that are most commonly reported in research | No chosen measure<br>• None of the  measures chosen for meta-analysis are reported in the results section of the study |
| Self-report data<br>A self-report measure of pain is applied to children aged three years or older, assessing the same children and same incident of pain as the observational measure | To ensure that there is a valid measure of self-report to compare against. The cut-off of three years is based on the recommendations for clinical use of self-report scales (Royal College of Nursing, 2009). | No self-report<br>• No self-report data reported<br>• Self-report not made independently<br>• Self-report scores reported from sample under years of age<br>• Self-report measures are applied to a different group of children than observational measure<br>• Self-report was applied at a different time point than observational measure |
| Correlation<br>The results of a correlation analysis between the observational measure and the self-report measure of pain is reported by the study | For the purposes of the meta-analysis a correlation was required as the assessment of association between observation and self-report scores of pain | No correlation<br>• Study does not report a correlation analysis between self-report and observational assessments of pain |

The initial search returned 18335 articles. The full results of the screening process are presented in Figure 1 in accordance with PRISMA guidelines for conducting and reporting systematic reviews and meta-analyses (Moher, Liberati, Tetzlaff & Altman, 2009). Initial removal of duplicates and papers which could not be found or accessed left a total of 9722 papers for title and abstract screening.

The most common reasons for exclusion at abstract screening were that the paper was not a quantitative research study (*n* = 2449), or that it used only self-report measures in the assessment of pain (*n* = 1013). The remaining 2244 articles had their methodology and results reviewed in more detail against the inclusion and exclusion criteria. Articles reporting a literature review or meta-analysis that used similar inclusion criteria to the current study, specifically regarding age ranges and use of observational measures, were screened for any papers that had been missed from the original search. Although three additional papers were identified through this process, none of them met inclusion criteria.

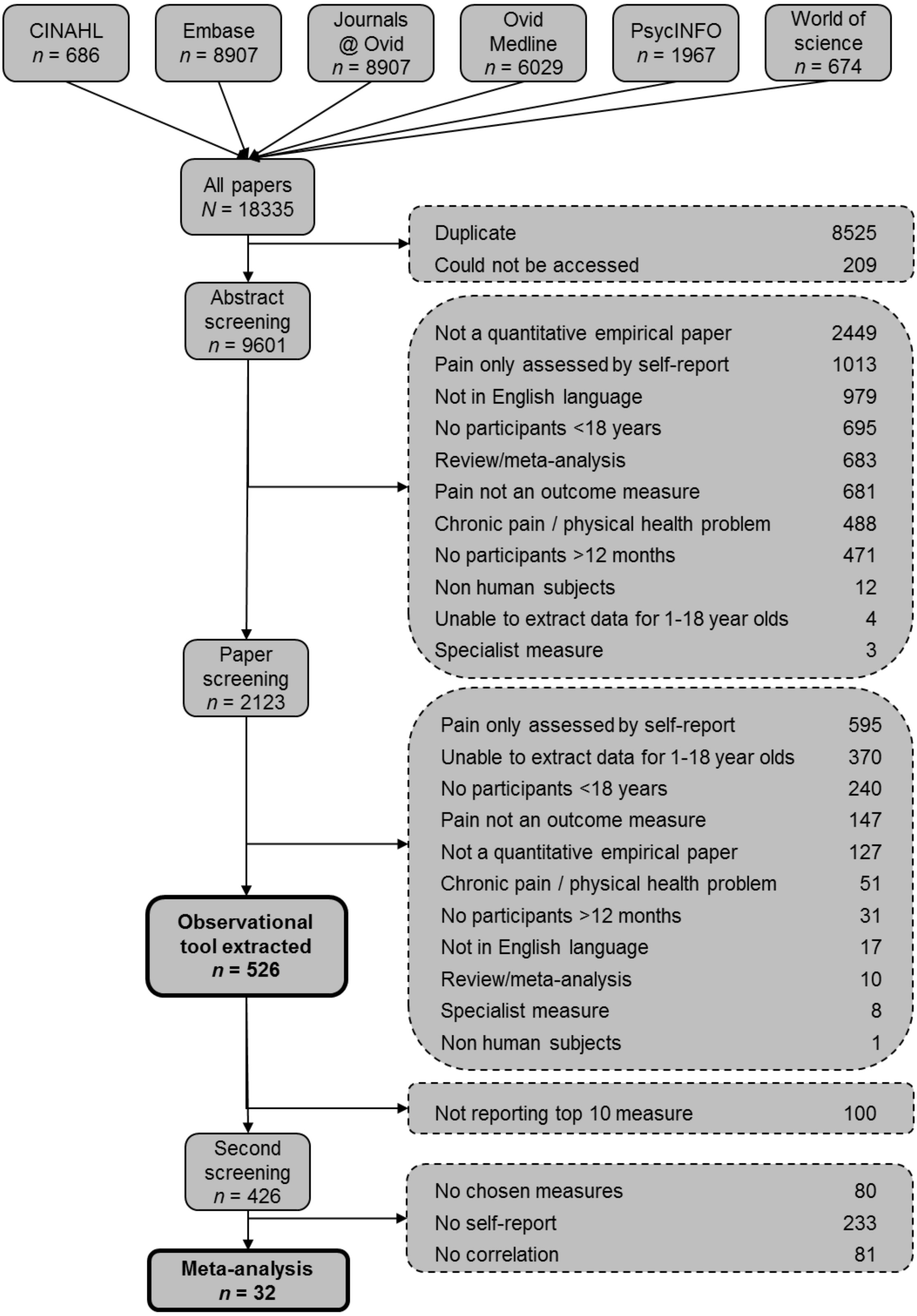

CINAHL
n = 686
Embase
n = 8907
Journals @ Ovid
n = 8907
Ovid Medline
n = 6029
PsycINFO
n = 1967
World of science
n = 674
All papers
N = 18335
Duplicate 8525
Could not be accessed 209
Abstract screening
n = 9601
Not a quantitative empirical paper 2449
Pain only assessed by self-report 1013
Not in English language 979
No participants <18 years 695
Review/meta-analysis 683
Pain not an outcome measure 681
Chronic pain / physical health problem 488
No participants >12 months 471
Non human subjects 12
Unable to extract data for 1-18 year olds 4
Specialist measure 3
Paper screening
n = 2123
Pain only assessed by self-report 595
Unable to extract data for 1-18 year olds 370
No participants <18 years 240
Pain not an outcome measure 147
Not a quantitative empirical paper 127
Chronic pain / physical health problem 51
No participants >12 months 31
Not in English language 17
Review/meta-analysis 10
Specialist measure 8
Non human subjects 1
Observational tool extracted
n = 526
Not reporting top 10 measure 100
Second screening
n = 426
No chosen measures 80
No self-report 233
No correlation 81
Meta-analysis
n = 32

*Figure 1*: Flow diagram illustrating screening process and reasons for exclusion

## 1.3.2. Meta-analysis

### 1.3.2.1. Quality rating.

The quality rating framework presented in Table 3 was developed to assess the quality and risk of bias within the literature. The framework assessed quality across five domains; sample selection, induction of pain, observer blinding, use of observer scale, and use of self-report scales. All domains were rated on a likert scale from 0 - 3 with 3 indicating the highest level of quality and methodological rigour. The quality rating score was obtained by summing the obtained values and dividing the sum by the maximum potential score of 15, thereby producing a quality score between zero and one, with one indicating a primary study which achieved ratings of high quality in all domains.

*Table 3*: Quality rating framework applied to primary studies in order to calculate quality index scores for quality weighted analysis

| **Domain** | **0 Unable to rate** | **1 Low quality** | **2 Moderate quality** | **3 High quality** |
|---|---|---|---|---|
| Sample selection | Clinic or setting not identified | Single restricted/non-random sample e.g. specialist clinic | Multiple restricted/non-random samples e.g. multiple clinics | Random or total population study |
| Cause of pain | N/A meets exclusion criteria | Cause of pain named or described, but multiple causes of pain/procedural variations grouped together in reporting correlation. | Specific painful procedure outlined, or multiple procedures with correlations separated by procedure. Multiple clinicians or teams administering procedure, or clinician/team not described. | Painful procedure clearly outlined and administered by a single clinician or team (where multiple clinicians are required for procedure). In cases of multiple procedures either all procedures administered by same clinician/team or distinct clinician/team per procedure as reported in results. |
| Risk of observer bias | No description of how rating was made / rater not identified. | Pain assessed by person with knowledge of experimental condition or self-report scores, or person administering procedure. | Pain assessed by rater(s) blinded to self-report scores and experimental condition (if appropriate). | Pain assessed by rater(s) blinded to self-report and experimental condition, in addition to at least one other assessment of pain by another rater. |
| Use of observer measure | N/A – meets exclusion criteria | Study does not describe any evidence supporting use of the tool as an observer measure. | Study describes evidence base for use of tool as observer measure but not pertaining to use in described sample. | Study clearly describes evidence justifying use of tool as an observer measure to assess pain in described sample. |
| Use of self-report | Self-report scores given with no description of how they were obtained. | Use of own measure, or measure described but not cited. | Study describes evidence base for use of tool as a self-report measure but not pertaining to use in described sample. | Study clearly describes evidence justifying use of self-report measure in described sample and/or describes procedure to endure children’s ability to understand and use measure. |

#### 1.3.2.2. Data extraction.

The results of reported correlation analyses between the self-report and observational rating tool were extracted from primary studies. A protocol for data extraction was developed to ensure consistent decisions were made regarding the choice of correlation values in studies where multiple correlations were reported. The reliability of data extraction was assessed using a 20% random sample. A research assistant independently extracted correlation values in accordance with the protocol and retrieved the same values as the author in all cases.  Where a primary study reported relevant correlations for multiple subsamples of different age groups containing unique participants, both were extracted and entered separately into the synthesis.

#### 1.3.2.3. Analysis strategy.

To meet the first aim of the current study, the methodology and results of all papers meeting the literature review inclusion criteria were reviewed to extract the observational measures used. The details of each named observational measure were recorded. The number of uses of each named measure was recorded to calculate which measures were reported most frequently in the literature. After the five most frequently used measures had been determined, all papers were reviewed using the meta-analysis inclusion criteria (Table 2), those meeting inclusion criteria underwent quality rating and data extraction. The top five measures were selected for further analysis because together they represented approximately two-thirds (66.57%) of the reviewed literature. Given that only a small portion of the reported studies were likely to meet the further inclusion criteria for the meta-analysis, it was thought unlikely there was a sufficient wealth of data for a meta-analytic approach to be worthwhile when applied to measures other than the five most commonly used.

The extracted correlation scores were recorded as Pearson's r correlation coefficients. Where studies reported using nonparametric approaches such as Spearmans Rho or Kendall's Tau, then the Pearson coefficient was approximated using the transformations reported by Rupinski and Dunlap (1996).

Meta-analyses were conducted using the random-effects model. The random-effects is used to estimate the mean of a distribution of possible correlations. It is accepted that true variation may occur due to the idiosyncratic characteristics of the individuals being studied or the unique circumstances of experimental procedure. In contrast to a fixed-effect model, small n studies are not discounted based on sample size alone, nor are large n studies necessarily given a larger weighting. Since each study provides information about the correlation between observation and self-report in a unique sample and a unique set of procedures, the goal of the synthesis is to capture all of the available effect sizes that describe these difference contexts, without allowing any single study to assert disproportionate influence over the final estimate. Given the variations in methodological procedures and in sample characteristics that were observed in the primary studies, it is highly unlikely that all the studies were measuring the same effect to the same precision. Therefore it was not appropriate to assume a common effect size across all studies, hence the use of a random-effects model rather than a fixed-effects model.

As an additional exploration of the effects of the quality of the included studies on the synthesis, the quality-effects model was also calculated. The quality effects model (Doi & Thalib, 2008) extends the random effects model by explicitly including rating of methodological quality in addition to the size of the sample in the estimation of precision. In this review the quality effects model was calculated using the total score from the quality framework outlined in Table 3. The quality effects model can be interpreted as the meta-analytic synthesis that would have been obtained had all of the studies been of the same methodological quality as the best study in the review. Accordingly, the quality effects model provides a measure of attrition attributable to methodological variation.

Within the calculation of the random-effects model, the DerSimonian and Laird method (DerSimonian & Laird, 1986) was applied to calculate the between studies variation (tau). Due to the variation in the reported methodology of the included studies, it was important to assess heterogeneity in the meta-analysis. A heterogeneous effect refers to distributions of effects in the primary studies that are too great to be idiosyncratic variation in the correlation between observation ratings of pain and self-report. These may reflect measurement error, individual differences within the sample, or methodological variation in study procedures. Cochrane's *Q* was also applied to the analysis of heterogeneity, it is a computation based upon the deviation of each effect size from the mean of all studies. If the *Q* value is significant at alpha < .01 then there is definite evidence of heterogeneity. In the current review the Higgins $I^2$ statistic was the primary calculation used to define the cut off for "problematic heterogeneity". Higher $I^2$ values indicate variation in effect that cannot be attributed to true variation in the distribution of effect in the population. The threshold for defining problematic heterogeneity in the current review was defined as a Higgins $I^2$ value greater than 75%. A high threshold for problematic heterogeneity was determined appropriate due to both the considerable variation observed in the methodologies of the primary studies, and the subjective nature of pain ratings.

In cases of problematic heterogeneity, a leave-one-out analysis was conducted to identify primary studies that exerted a disproportionately influential effect on the meta-analytic synthesis. If omitting a study resulted in an effect outside of the 95% CI for the complete meta-analysis then that study was deemed to have a disproportionate influence. The study was excluded and the synthesis was repeated.

Subgroup analyses were applied to all meta-analyses with a sufficient number of primary studies, to attempt to explore potential sources of heterogeneity across the studies. Differences in the weighted estimates between groups of studies was calculated using Cochrane's *Q*.

Publication bias and small study effects were explored through visual and statistical inspection of the funnel plot. A funnel plot is a scatterplot of the observed effects from each of the primary studies against a measure of study precision. A funnel plot provides a method of detecting systematic heterogeneity. The assumption is that studies with high precision will be plotted near the average provided by the meta-analytic synthesis, whereas studies with low precision will be spread evenly on both sides of the average, creating a roughly funnel-shaped distribution. Studies that fall outside of the desired 'funnel' suggest the presence of publication bias in the group of primary studies. A trim and fill procedure (Duval & Tweedle, 2000a; Duval & Tweedle, 2000b) was applied to any meta-analysis in which publication bias was identified. The trim and fill procedure builds on the assumption that publication bias would lead to an asymmetrical funnel plot. An iterative algorithm was applied to the data to remove the most extreme small studies from the side of the funnel plot, re-computing the effect size at each iteration until the funnel plot is symmetric. In theory, the trim and fill procedure yields an unbiased estimate of the effect size. However, the trim and fill procedure also reduces the variance of the effects, yielding a too narrow confidence interval. Therefore, the algorithm then adds the original studies back into the analysis and imputes a mirror image for each. Because the funnel plot method is based on an assumption of normal distribution of data points, it was not applied where the synthesis revealed unacceptable levels of heterogeneity.

An additional exploration of the effects of publication bias was achieved through the calculation of the fail-safe N (Rosenthal, 1979). The fail-safe N provides an estimation of how many non-significant studies would be required for the observed effect to no longer be significant. A large fail-safe N suggests that the omnibus test can be considered robust to the effects of publication bias.

## 1.4. Results

### 1.4.1. Observational Assessments of Pain in the Literature

To address the first aim of the study and explore the use of observational tools reported in the research literature, a systematic search of six databases was conducted, and reported observational tools were extracted.

The measures used in the 526 papers that met inclusion criteria were recorded and are presented in Table 4. Variations of measures such as the Modified Children's Hospital of East Ontario Pain Scale (mCHEOPS; Splinter, Semelhago, & Chou, 1994) were grouped with the original measure where the original measure was clearly identifiable or named. In the case of the various different faces scales, faces scales were reported separately due to the use of different graphics or images for the faces.

Across the 526 papers reviewed, a total of 62 observational measures were identified. However, 46 of the identified measures were reported in fewer than ten papers each, and 29 were reported in only one paper each. Thirteen papers were reviewed which did not reference an observational measure, but instead described an observation based assessment designed specifically for the purposes of their study. The category of "Study specific measure" was not used if the measure was also found referenced in other reviewed papers i.e. if it was the first paper to publish a measure which went on to be more widely used. Only eight measures could be found that were reported in more than 20 papers each, and three of those eight may be better referred to as techniques rather than as published measures. The Visual Analogue Scale (VAS), Numerical Rating Scale (NRS), and Categorical Rating Scale

(CRS) are all variations of global rating scales which use linear, numeric scales or categories such as “mild/moderate/severe” to rate pain. Although there are some published measures which incorporate these rating techniques, for the purposes of the current review, a paper was categorised as using a VAS, NRS or CRS if the technique was used in isolation with no guidance as to which behaviours to use to determine the final rating.

The five most commonly reported measures were the Children’s Hospital East Ontario Pain Scale (CHEOPS; McGrath et al., 1985), the Visual Analogue Scale (VAS; Price, McGrath, Rafii & Buckingham, 1983), the Face, Legs, Activity, Crying and Consolability (FLACC; Voepel-Lewis, Shayevitz & Malviya, 1997), the Observer Pain Scale (OPS; Hannallah et al., 1987), and the Wong-Baker Faces Scale (WBF; Wong & Baker, 1988). The majority of the 526 reviewed papers reported using one or more measures from the five most commonly reported measures (*k*[1] = 346). Of the 346 papers which reported using one of the five most common measures, only 32 met the further inclusion criteria (see Table 3) required to be included in the meta-analytic stage of the review. Further detail on the measures and on the 32 primary studies reporting an association between those measures and self-report of pain can found in the meta-analyses presented in section 1.3.3.

[1] *k* is used to denote number of published studies

*Table 4*: Measures extracted from final paper sample with frequency (*k*) of use in the literature. Frequency totals more than 526 due to use of multiple measures in many of the studies reviewed.

| **Name of measure** [2] | **Authors** | ***k*** |
|---|---|---|
| Children's Hospital of East Ontario Pain Scale | McGrath et al. (1985) | 121 |
| Visual Analogue Scale | Multiple versions described | 87 |
| Face Legs Activity Crying Consolobility Scale | Voepel-Lewis et al. (1997) | 81 |
| Objective Pain Scale | Broadman et al. (1988) | 71 |
| Wong & Baker FACES scale | Wong & Baker (1988) | 26 |
| Children's and Infants' Postoperative Pain Scale | Büttner & Finke (2000) | 24 |
| Categorical Rating Scale | Multiple versions described | 24 |
| Numerical Rating Scale | Multiple versions described | 22 |
| Parents Postoperative Pain Measure | Chambers et al. (1996) | 18 |
| Toddler Preschooler Postoperative Pain Scale | Tarbell et al. (1992) | 15 |
| Faces Pain Scale | Bieri et al. (1990) | 13 |
| Study specific pain assessment described without measure being named or cited | | 13 |
| Sound, Eye and Motor scale | Wright et al. (1991) | 12 |
| Modified Behaviour Pain Scale | Taddio et al. (1995) | 11 |
| Child Facial Coding System | Gilbert et al. (1999) | 10 |
| Faces Pain Scale-Revised | Hicks et al. (2001) | 10 |
| Maunuksela Behavioural Pain Scale | Maunuksela et al. (1987) | 10 |
| Non Communicating Child Pain Checklist | Breau et al. (2002) | 9 |
| Colour Analogue Scale | McGrath et al. (1996) | 4 |
| The COMFORT scale | Ambuel et al. (1992) | 4 |
| Maunuksela faces VAS pain scale | Maunuksela et al. (1987) | 4 |
| Procedural Behaviour Checklist | LeBaron & Zeltzer (1984) | 4 |
| Behavioural Observational Pain Scale | Hesselgard et al. (2007) | 3 |
| Behaviour Checklist | Goodenough et al. (1997) | 3 |
| Preverbal, early verbal pediatric pain scale | Schultz et al (1999) | 3 |
| University of Wisconson Childrens Hospital Pain Scale | Soetenga et al. (1999) | 3 |
| Dalhousie Everyday Pain Scale | Fearon et al. (1996) | 2 |
| Derbyshire Children's Hospital Pain Tool | Unpublished | 2 |
| Individualised Numeric Rating Scale | Solodiuk & Curley (2003) | 2 |
| Kuttner & LePage FACES scale | Kuttner & LePage (1989) | 2 |
| Nurses Assessment of Pain Index | Stevens (1990) | 2 |

[2] Name is based on most commonly used name for measure found in the review, name has been omitted in cases where measure was referred to by reference only.

| **Name of measure** [2] | **Authors** | ***k*** |
|---|---|---|
| Objective Pain Discomfort Score | Steward (1975) | 2 |
| Pediatric Pain Profile | Hunt et al. (2004) | 2 |
| Total Quality Pain Management | Foster & Varni (2002) | 2 |
| Alder Hey triage pain score | Stewart et al. (2004) | 1 |
| AIIM Pain Discomfort Scale | Brown & Fisk (1992) | 1 |
| Baby FACS | Oster & Rosenstain (1993) | 1 |
| Behavioural Pain Scale | Payen et al. (2001) | 1 |
| Measure not named | Cameron et al. (1992) | 1 |
| Child Pain Scale | Gauvain-Piquard et al (1987) | 1 |
| Echelle Douleur Enfant San Salvador | Collignon & Giusiano (2001) | 1 |
| Facial Action Coding System | Ekamn & Friesan (1978) | 1 |
| Facial Affective Scale | McGrath et al. (1996) | 1 |
| Global Assessment of Behavioural Reaction | Juniper et al. (1991) | 1 |
| Observational Pain-Discomfort Scale | Buttner et al. (1990) | 1 |
| Hester Poker Chip Tool | Hester (1979) | 1 |
| Izard's Coding for Facial Signs of Pain | Rowland et al. (1989) | 1 |
| KKU Pediatric Pain Assessment Tool | Jongudomkarn et al. (2008) | 1 |
| Measure not named | Krane et al. (1987) | 1 |
| Multidimensional Assessment of Pain Scale | Ramelet et al. (2007) | 1 |
| Modified Pediatric Observer Pain Scale | Wolf et al. (1990) | 1 |
| Multiple Size Poker Chip Tool | St-Laurent-Gagnon et al. (1999) | 1 |
| Neonatal Facial Coding System | Grunau et al. (1990) | 1 |
| Pain Behaviour Checklist | Peters (2007) | 1 |
| Pain Indicator for Communicatively Impaired Children | Stallard et al. (2002) | 1 |
| Princess Margaret Hospital Pain Assessment Tool | Robertson (1993) | 1 |
| Post Operative Pain Score | Attia et al. (1987) | 1 |
| Pain Rating Scale | Joyce et al. (1994) | 1 |
| Royal College of Emergency Medicine Composite Pain Tool | Royal College of Emergency Medicine (2004) | 1 |
| Riley Infant Pain Scale | Schade et al. (1996) | 1 |
| Toddler Discomfort Index | Tomlinson & Stewart (2008) | 1 |
| Universal Pain Assessment Tool | University of California (2004) | 1 |
| Verbal Pain Score | Güleç et al. (1998) | 1 |

### 1.4.2. Selection of Measures for Meta-analysis

Papers including any of the ten most commonly reported observational measures of pain were screened for inclusion of any self-report measures of pain, and the reporting of a correlation analysis between self-report and observational pain scores[3]. The results in Table 5 show that few data are available pertaining to the relationship between self-report and many of the observational measures. The top five most commonly reported measures were selected for meta-analysis because the results of the second screening suggest these measures presented sufficient data for a worthwhile synthesis. Although only one study could be found which reported on a correlation between self-report and the Objective Pain Scale (OPS), it is reported on below due to the frequency with which the OPS is cited in the literature.

*Table 5:* The number of papers reporting the use of one of the ten most commonly used observational pain assessments, self-report of pain, and a correlation between observation and self-reported pain ratings

| **Measure** | **Number of studies** | **Self-report included** | **Correlation reported** |
|---|---|---|---|
| Children's Hospital of East Ontario Pain Scale | 121 | 31 | 7 |
| Visual Analogue Scale | 87 | 48 | 15 |
| Face Legs Activity Crying Consolobility Scale | 81 | 25 | 9 |
| Objective Pain Scale | 71 | 8 | 1 |
| Wong & Baker FACES scale | 26 | 20 | 7 |
| Children's and Infants' Postoperative Pain Scale | 24 | 1 | 0 |
| Categorical Rating Scale | 24 | 11 | 2 |
| Numerical Rating Scale | 22 | 11 | 3 |
| Parents Postoperative Pain Measure | 18 | 8 | 4 |
| Toddler Preschooler Postoperative Pain Scale | 15 | 1 | 0 |

### 1.4.2. Quality Assessment of the Most Commonly Used Measures

To address the first aim of the review and provide further assessment of the use of observational pain assessments in the literature, the quality of the 32 primary

[3] At this stage of screening, the specific criteria listed in the 'meta-analysis full review' section of table 2 were not applied, therefore the numbers reported in table 5 do not align with the number of primary studies included in the final meta-analyses.

studies utilising the CHEOPS, VAS, FLACC, OPS and WBF were explored in more detail. Quality was assessed across primary papers according to the quality framework presented above (Table 3).

#### 1.4.2.1. Sample.

Five studies failed to report any information regarding the study setting or how the sample was obtained. A single point of recruitment was reported in 20 studies, and seven studies reported multiple recruitment sites. The largest selection of recruitment sites was reported by Boivin et al. (2008) who reported recruitment from 25 GPs in the Lorraine region of France.

#### 1.4.2.2. Cause of pain.

Many of the primary studies were intervention studies examining the effects of variations in procedure, such as changes to technique, or the effect of different analgesics. For the purposes of this review such variations were considered “multiple procedures” in the quality rating. The majority of studies were given a low quality rating (*k* = 21) because, although they often separated out the key outcome results by procedure, they grouped all procedures together in the reporting of the correlation between observer measure and self-report, meaning that the pain scores were related to multiple different painful procedures. Eight studies received a moderate quality rating, reporting specific procedures but either reporting that the procedure was carried out by multiple professionals, or failing to describe those responsible for the painful procedure. Three studies received a high quality rating, with the source of pain being a single, well described procedure, administered by the same person or team to each participant.

#### 1.4.2.3. Risk of observer bias.

The majority of studies received a low quality rating for failing to blind observers (*k* = 15). Many studies reported blinding to experimental groups, however there was frequently ambiguity regarding whether observers had been blinded to child ratings of pain, which may influence the observer rating and so was rated as low quality for this domain. Six studies clearly reported blinding observers from self-report ratings and experimental groupings, and a further seven reported the addition of at least one other rating by another observer, further reducing the potential risk of bias. Four studies were categorised as “unable to rate” in this domain, as there was no procedure reported for obtaining observer scores beyond the naming of the observational tool, therefore it was impossible to determine to what extent observers had been blinded. Chadha et al (2013) was rated twice for this domain as the procedure differed according to the two measures used. In the case of the FLACC, which was completed by an observer, the study received a high quality rating, as blinding procedures were clearly described. However, in the case of the WBF, which was completed by a parent, there was nothing described in the procedure to suggest that parents had been blinded to the self-report scores provided by their children. Therefore, in applying the WBF Chadha et al. received a low quality rating.

#### 1.4.2.4. Use of observation scales.

Six studies received a high quality rating for their use of observation scales, describing evidence that explicitly supported the use of the named tool in the sample reported by the primary study. However, the majority of studies received a rating of moderate quality (*k* = 12) or low quality (*k* = 14). Poorest ratings in this domain were obtained by studies reporting use of global rating scales. Studies reporting the CHEOPS and FLACC fulfil criteria for a moderate quality rating simply by citing the measure, as the measures were originally published as observational tools. However, references for the VAS were rarely given at all, and the original validation of the WBF does not provide support for its use as an observational tool. Only two of the fourteen studies using the VAS as an observational tool made explicit reference to an evidence base supporting its validity when used in this way (Kelly et al., 2002; Bearden et al., 2012).

Several of the 32 studies reported multiple observation measures, and so because ratings in this domain are based on appropriateness of the scale to the described sample, a separate rating was given based on the appropriateness of each relevant measure. Bringuier et al. (2009) received a high quality rating for the use of the CHEOPS, but only moderate quality ratings for the FLACC and OPS. Chadha et al. (2013) received a moderate quality rating for the FLACC, but a low quality rating for the WBF. Risaw et al. (2017) received a high quality rating for their use of the FLACC, but a low quality rating for WBF.

#### 1.4.2.5. Use of self-report scales.

Use of self-report scales was the highest rated of the five domains, with 18 studies receiving a high quality rating. Many studies provided justification for the choice of self-report with their sample. Some studies with wide age ranges used multiple self-report scales to ensure that appropriate scales were available for all participants. Studies frequently included screening procedures assessing comprehension of the self-report scales, or explicitly described teaching procedures to ensure that children understood the scales and could provide meaningful ratings. Only three studies failed to provide evidence to support their choice of self-report scale, typically because they used their own self-report scale that had no published evidence base. Eleven studies described some evidence supporting the choice of self-report scale but either failed to mention the sample used in the supporting studies, or explicitly described studies with samples that did not match the sample reported in the primary study.

The rating of quality may reflect fundamental limitations in the measurement of subjective experience; however, all of the reviewed studies contain at least one area of potential contamination and bias on a wide range of criteria and the overall quality of this corpus of evidence should be considered as poor.

### 1.4.3. Associations Between Observational Assessments and Self-report

To address the second aim of the study exploring the association between the five most commonly used observational assessments and self-report measures of pain, a series of meta-analyses were undertaken, synthesising the available literature where studies reported a correlation between one of the five most frequently used behavioural assessments and self-report measures of pain. These meta-analyses incorporated the ratings of study quality, as described above.

#### 1.4.3.1. Children’s Hospital East Ontario Pain Scale.

The Children’s Hospital East Ontario Pain Scale (CHEOPS; McGrath et al., 1985) is a behavioural tool which requires observers to identify and rate six behaviours. It has the unusual score range of 4 - 13, with scores under six indicating no pain. A modified version of the CHEOPS is also available which simplifies the scale to a 0 - 2 rating of five behaviours, resulting in a total score in the 0 - 10 range (Splinter, Semelhago, & Chou, 1994).

The primary studies included in the analysis are reported in Table 6. There were eight primary studies reporting a total of *N* = 517 participants. The analysis included participants from the age of four to 15 years, with the majority of participants being under ten. The majority of studies in this analysis reported using variations of the faces scale for self-report (WBF, Wong & Baker, 1988; Faces Pain Scale Revised; FPSR, Hicks, von Baeyer, Spafford, van Korlaar & Goodenough, 2001; Faces Pain Scale, Bieri, Reeve, Champion, Addicoat & Ziegler, 1990), with two studies reporting using the VAS, and one using the Oucher (Beyer, Denyes & Villarruel, 1992). All of the studies included in this meta-analysis reported using the original CHEOPS as described by McGrath et al. (1985).

*Table 6*: Methodological details and quality rating of primary studies reporting a correlation between pain scores obtained from CHEOPS and self-report. Details presented as reported in published paper.

| Study | *n* [4] | Age range (years) | Sample | Pain | Blinding | Observation | Self-report | Source of Pain | CHEOPS rater(s) | Self-report scale | *r* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Beyer et al. (1990) | 8 | 3-7 | 0 | 1 | 2 | 3 | 3 | Surgery | Nurse | Oucher | .47 |
| Bringuier et al (2009) | 19 | 4-7 | 1 | 1 | 3 | 3 | 2 | Surgery | Nurse/ anaesthetist | FPSR | .48 |
| Cassidy et al (2002) | 58 | 5 | 2 | 1 | 2 | 3 | 3 | Vaccination | Blinded rater | Faces | .49 |
| Hee et al. (2003) | 120 | 8-15 | 1 | 1 | 2 | 2 | 3 | Cannulation | Nurse/ anaesthetist | VAS | .21 |
| Lee & White-Traut (1996) | 126 | 3-7 | 1 | 1 | 1 | 2 | 2 | Venipuncture | Not named | WBF | .47 |
| Sikorova & Hrazdilova (2011) | 60 | 5-10 | 1 | 1 | 0 | 2 | 2 | Venipuncture | Researcher | WBF | .62 |
| Tyler et al. (1992; 3-6.5yrs) | 16 | 3-6.5 | 2 | 2 | 0 | 2 | 3 | Surgery | Investigator | Faces | .74 |
| Tyler et al. (1992; 6.5-12yrs) | 10 | 6.5-12 | 1 | 1 | 0 | 2 | 2 | Surgery | Investigator | Faces | .74 |
| Vessey et al. (1994) | 100 | 3.5-12.9 | 1 | 1 | 1 | 2 | 3 | Venipuncture | Research assistant | WBF | .63 |

[4] *n* refers to the sample reported for the correlation that was extracted from the primary study. If a specific *n* value was not reported for the extracted correlation, then it was assumed that the correlation was based on the whole study sample.

The random effects model reported in Figure 2 estimated a weighted average correlation of *r* = .52, 95% confidence interval (CI) [.39, .64]. This suggests a moderate positive correlation between self-report of pain and observation using the CHEOPS. The level of heterogeneity in the portions reported in the primary studies was found to be within acceptable parameters for the current review (tau$^2$ = .04, Higgin's I$^2$ = 65.2%; *Q* = 23.00, *p* = .003).

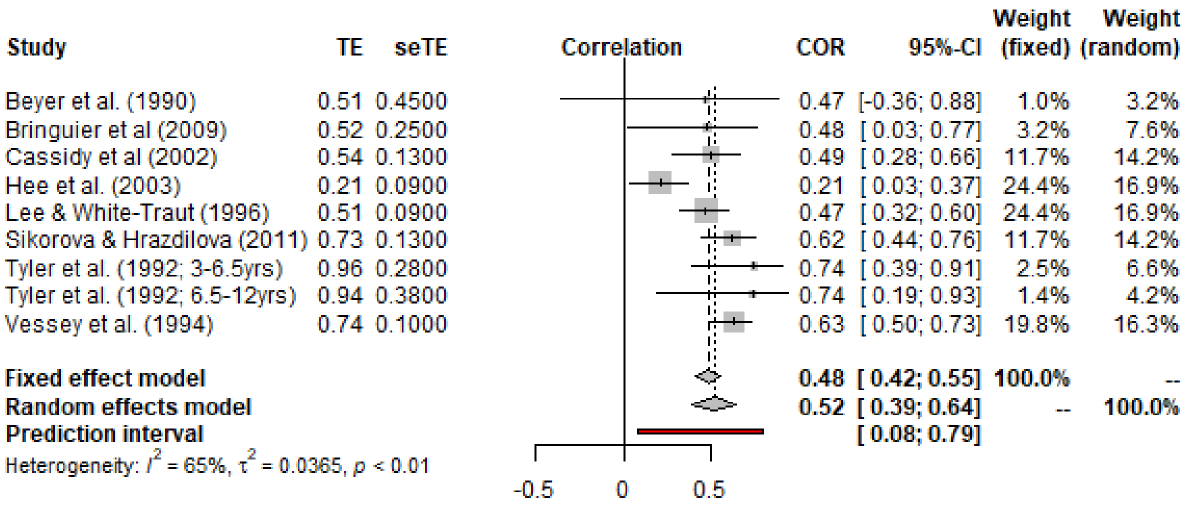


*Figure 2*: Forest plot illustrating the meta-analytic synthesis of correlations between pain ratings obtained by CHEOPS and self-report. TE = Measure of the effect, transformed into *z* score; seTE = measure of standard error; COR = *r* score.

The quality effect model estimated a weighted average correlation of *r* = .54 95% CI [.38, .66]. The quality effects model evidences a 2% increase relative to the random effects estimate. Accordingly, when the synthesis includes information about the methodological quality of the studies there is no substantive change in the estimation of the weighted average correlation.

As can be seen from Figure 3 the funnel plot shows asymmetry in the published studies. A trim and fill procedure imputed one additional result to adjust the symmetry of the funnel plot. The uncorrected estimate of the effect size was *r* = .52,   95% CI [.39, .64], the adjusted estimate was *r* = .50, 95% CI [.37, .62]. The adjusted point estimate suggests a lower effect than the original analysis. The Rosenthal algorithm suggests a failsafe number of 426 unpublished null effect studies required to reduce the meta-analytic effect of the nine results reported here to a non-significant finding.

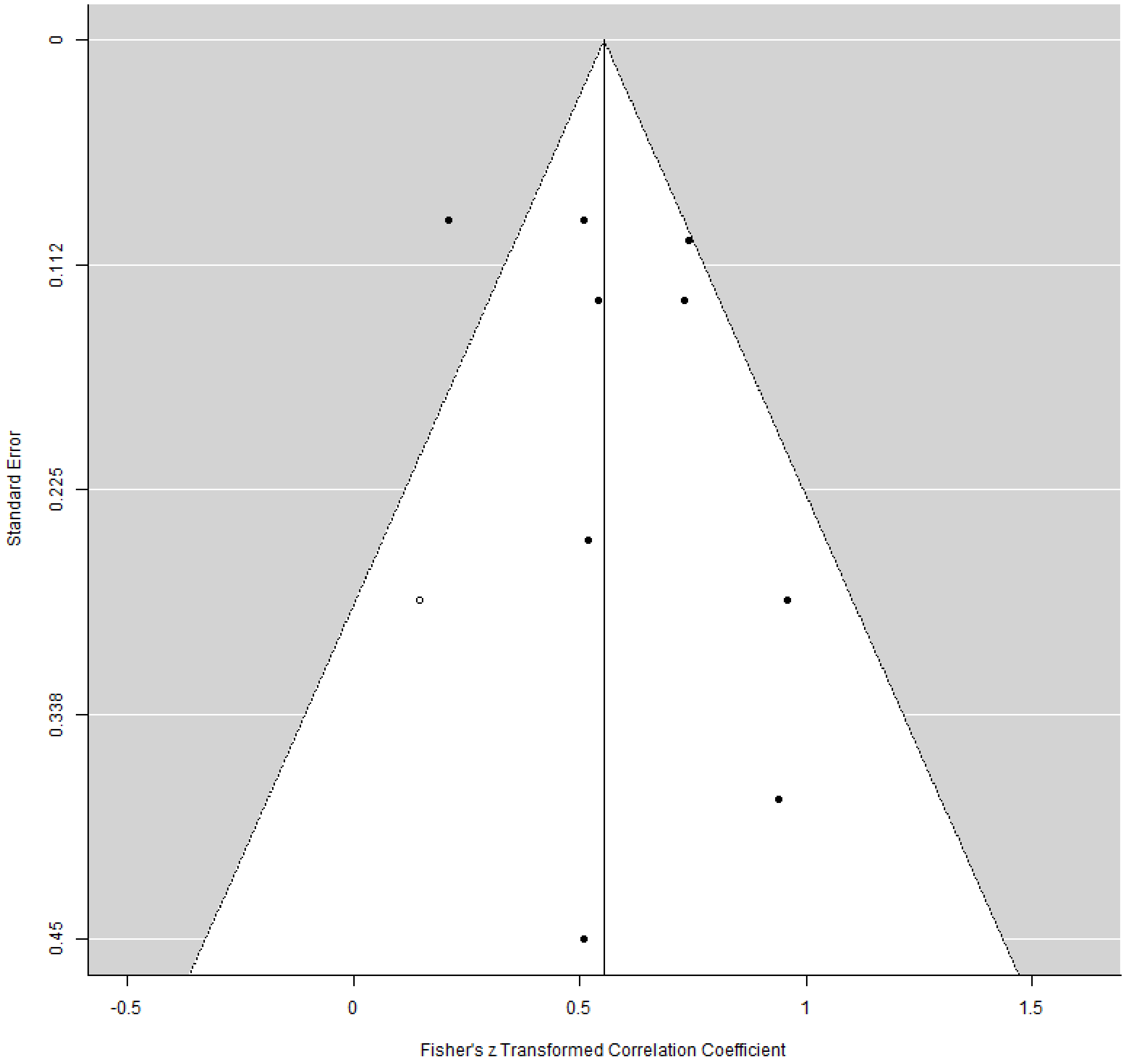


*Figure 3*: Funnel plot illustrating trim and fill procedure for CHEOPS analysis. Black dots indicate primary studies, white dots indicate studies imputed by trim and fill procedure.

To further explore the impact of uncontrolled covariates upon the correlation between self-report of pain and CHEOPS scores, a series of subgroup analyses were conducted. The first set of subgroup analyses exploring results according to quality rating found no significant differences in the synthesised *r* scores produced by studies of low, moderate, or high quality or those that could not be rated, regardless of domain. The results of these subgroup analyses are presented in Table 7.

*Table 7*: Subgroup analyses of primary studies reporting CHEOPS grouped by quality rating for each quality domain.

| | **Not able to rate (*k*)** | **Low quality (*k*)** | **Moderate quality (*k*)** | **High quality (*k*)** | ***Q*** | ***p*** |
|---|---|---|---|---|---|---|
| Sample selection | .47<br>(1) | .56<br>(6) | .48<br>(2) | -<br>(0) | 0.52 | .771 |
| Cause of pain | -<br>(0) | .54<br>(1) | .47<br>(8) | -<br>(0) | 0.44 | .508 |
| Use of blinding | .52<br>(3) | .63<br>(2) | .31<br>(3) | .48<br>(1) | 5.65 | .130 |
| Use of observer measure | -<br>(0) | -<br>(0) | .55<br>(6) | .49<br>(3) | 0.22 | .636 |
| Use of self-report | -<br>(0) | -<br>(0) | .64<br>(3) | .49<br>(6) | 1.30 | .254 |

The second set of subgroup analyses explored the influence of additional methodological variables. Although no difference was found when studies were analysed according to the source of pain described, a significant difference was found when studies were analysed according to the role of the person completing the CHEOPS, with researchers achieving significantly higher correlations to self-report than clinicians (*p* < .001). Table 8 presents the results of these analyses.

*Table 8*: Subgroup analyses of primary studies reporting CHEOPS grouped by type of pain rated and role of person rating CHEOPS.

| **Type of Pain** | **Acute Procedural (*k* = 5)** | **Post-Surgical (*k* = 4)** | | ***Q*** | ***p*** |
|---|---|---|---|---|---|
| | .49 | .62 | | 1.02 | .313 |
| **CHEOPS rater** | **Clinician (*k* = 3)** | **Researcher (*k* = 5)** | **Not Stated (*k* = 1)** | | |
| | .25 | .61 | .47 | 18.32 | < .001 |

### 1.3.4.2. Visual Analogue Scale.

The Visual Analogue Scale (VAS) is a commonly used assessment technique utilising a horizontal or vertical line with two anchor points on which the respondent marks along the line to indicate the level of pain they believe the child to be in. The distance from the bottom anchor point is then measured and reported as a score of pain. The VAS has been validated for use as a self-report scale in children six years and above (Von Bayer, 2006). Studies in the current review varied in the length of line used for a VAS, though 100mm was most typical, they also varied in the wording used for the two anchor points.

The primary studies included in the analysis are reported in Table 9. There were 14 studies reporting a total of *N* = 1188 participants. The analysis included participants from the age of three to 18 years. The majority of studies in this analysis report using the same VAS for self-report as used for observer assessment (*k* = 9). Three studies reported multiple self-report scales to account for the needs of different age groups in their study, however individual correlations for these measures against the VAS were not given.

*Table 9*: Methodological details and quality rating of primary studies reporting a correlation between pain scores obtained from VAS and self-report. Details presented as reported in published paper.

| **Study** | ***n*** | **Age range (years)** | **Sample** | **Pain** | **Blinding** | **Observation** | **Self-report** | **Source of Pain** | **CHEOPS rater(s)** | **Self-report scale** | ***r*** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Bearden et al. (2012) | 88 | 4-6 | 1 | 2 | 3 | 2 | 3 | Vaccination | Nurse | Child Anxiety & Pain Scale | .46 |
| Benini et al. (2004) | 16 | 7-18 | 1 | 3 | 1 | 1 | 3 | Venipuncture | Parent & observer | VAS | .18 |
| Boivin et al. (2008) | 239 | 4-12 | 2 | 1 | 1 | 1 | 2 | Vaccination | GP | FPSR / VAS | .82 |
| Breau et al. (2001) | 123 | 4.3-6.6 | 2 | 2 | 1 | 1 | 3 | Vaccination | Medical Technician | FPS | .60 |
| Cohen et al. (2004) | 39 | 8.8-11.1 | 1 | 3 | 1 | 1 | 2 | Vaccination | Nurse | VAS | .42 |
| Foster & Varni (2002) | 50 | 8-12 | 2 | 1 | 2 | 1 | 3 | Surgery | Parent | VAS | .75 |
| Goodenough et al (1999) | 110 | 3-15 | 1 | 2 | 3 | 1 | 3 | Venipuncture | Parent | VAS / WBF | .56 |
| Jensen (2012) | 100 | 3-12 | 2 | 1 | 2 | 1 | 2 | Dental extraction | Parent | WBF | .79 |
| Jylli & Olsson (1995) | 96 | 3-16 | 1 | 1 | 0 | 1 | 1 | Painful procedures | Parent | Smiley scale / VAS | .33 |
| Kelly et al. (2002) | 78 | 8-15 | 1 | 1 | 2 | 3 | 3 | Painful conditions including trauma | Parent | VAS | .63 |
| Knutsson et al. (2006) | 100 | 3-9 | 0 | 1 | 3 | 1 | 3 | Adenoidectomy | Nurse | WBF | .62 |
| Lamontagne et al. (1991) | 13 | 8-18 | 0 | 1 | 3 | 1 | 3 | Surgery | Physician | VAS | .59 |
| Singer et al. (2002) | 63 | 4-7 | 1 | 1 | 3 | 1 | 2 | Acute painful condition / procedure | Medical Practitioner | Smiley scale | .54 |
| Tan & Stafford (1992) | 73 | 5-16 | 0 | 2 | 1 | 1 | 1 | Laser treatment | Physician | VAS | .77 |

The random effects model reported in Figure 4 suggested a weighted average correlation of $r$ = 0.62, 95% CI  [.51,  .70]. This suggests a moderate positive correlation between self-report of pain and observation using the VAS. An unacceptable level of heterogeneity in the portions reported in the primary studies was observed (tau$^2$ = .07, Higgin’s I$^2$ = 84.4%; $Q$ = 83.54, $p$ < .001). This suggests that the estimates of the primary studies are biased by the presence of uncontrolled or confounding factors. The quality effect model reported a synthesis of $r$ = .60, 95% CI [.48, .69]. The quality effects model evidences an approximately 3.12% decrease relative to the random effects estimate. Accordingly, when the synthesis includes information about the methodological quality of the studies there is no substantive change in the weighted average correlations obtained.

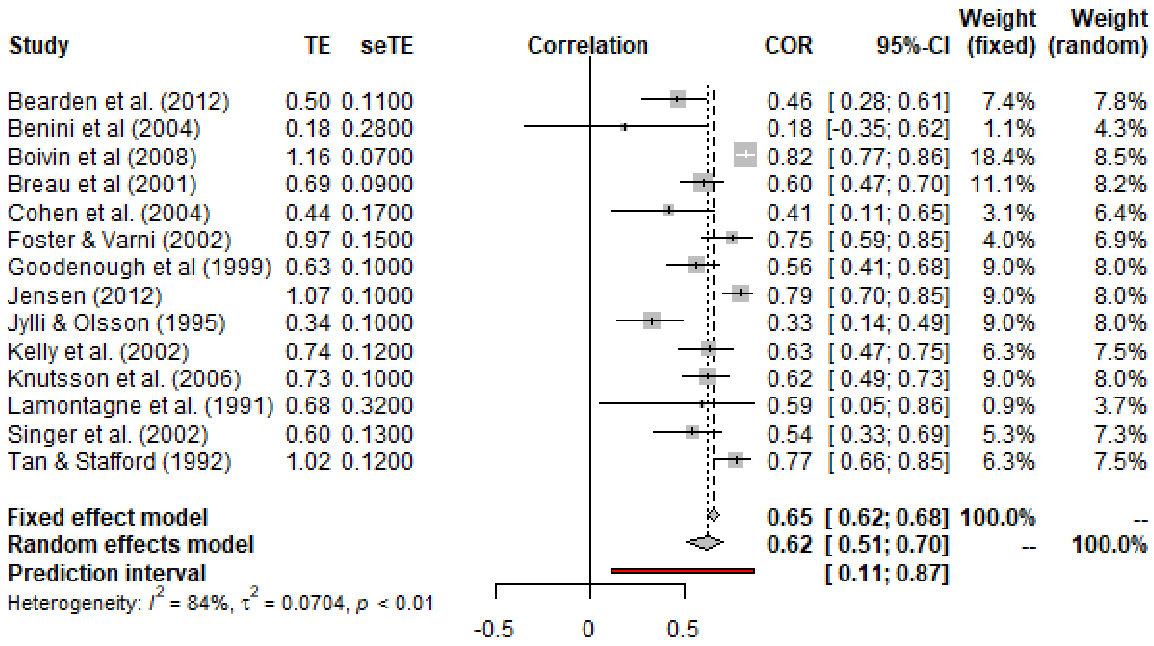


*Figure 4*: Forest plot illustrating the meta-analytic synthesis of correlations between pain ratings obtained by VAS and self-report. TE = Measure of the effect, transformed into *z* score; seTE = measure of standard error; COR = *r* score.

None of the studies met the criterion for removal therefore no corrections were made to the analysis based on the leave one out analysis. Because of the high levels of heterogeneity identified within the synthesis, a funnel plot was not considered appropriate.

The subgroup analyses for study quality found significant differences in the synthesised correlations when grouping by sample selection procedures, and when grouped according to their use of blinding, however, as seen in the results presented in Table 10, the direction of the effect is unclear across the groups. In the analysis of differences grouped by sample selection, the lowest correlation values were found in studies that were rated as low quality. When studies were grouped according to blinding procedures the lowest correlations were found in the groups which were categorised as “unable to rate”, however, this category comprised of only one study, and no clear pattern can be seen between the differences in correlations found between low, moderate and high quality studies.

*Table 10*: Subgroup analyses of primary studies reporting VAS grouped by quality rating for each quality domain.

| | **Unable to rate (*k*)** | **Low quality (*k*)** | **Moderate quality (*k*)** | **High quality (*k*)** | ***Q*** | ***p*** |
|---|---|---|---|---|---|---|
| Sample selection | .69<br>(3) | .48<br>(7) | .75<br>(4) | -<br>(0) | 14.29 | < .001 |
| Cause of pain | -<br>(0) | .66<br>(8) | .61<br>(4) | .35<br>(2) | 5.54 | .063 |
| Use of blinding | .33<br>(1) | .64<br>(5) | .73<br>(3) | .55<br>(5) | 17.20 | < .001 |
| Use of observer measure | -<br>(0) | .63<br>(12) | .46<br>(1) | .63<br>(1) | 3.25 | .197 |
| Use of self-report | -<br>(0) | .59<br>(2) | .69<br>(4) | .59<br>(8) | 0.97 | .616 |

To further explore the impact of uncontrolled covariates upon the association between observer pain scores obtained using the VAS and self-report pain scores a series of subgroup analysis were conducted. As with the other meta-analyses, type of pain, and the identity of the observer were conducted as subgroup analyses. The VAS was the only measure included in the meta-analyses where there were studies in which the self-report and observer scale were the same alongside studies in which the self-report scale differed, allowing a sub-group analysis exploring the effects of using the same scale for observers and self-report. The analysis, presented in Table

11, found no difference between studies comparing observational VAS to self-report VAS and those comparing observational VAS to different self-report measures.

*Table 11*: Subgroup analyses of primary studies reporting CHEOPS grouped by type of pain rated, role of person rating VAS, and type of self-report scale used.

| **Type of Pain** | **Acute Procedural (*k* = 8)** | **Post-Surgical (*k* = 3)** | **Other (*k* = 3)** | ***Q*** | ***p*** |
|---|---|---|---|---|---|
| | .64 | .66 | .50 | 2.73 | .256 |
| **VAS rater** | **Clinician (*k* = 8)** | **Parent (*k* = 5)** | **Not Stated (*k* = 1)** | | |
| | .63 | .63 | .18 | 3.72 | .156 |
| **Self-report scale** | **VAS (*k* = 6)** | **Other (*k* = 5)** | **VAS & Other (*k* = 3)** | | |
| | .62 | .62 | .61 | 0.00 | .999 |

### 1.3.4.3. Face, Legs, Activity, Crying and Consolability.

The Face, Legs, Activity, Crying and Consolability (FLACC; Voepel-Lewis, Shayevitz & Malviya, 1997) rates the five behavioural domains forming the name of the measure, each on a 0 - 2 scale, to produce a score of pain intensity between 0 - 10. It has been validated in the assessment of both brief procedural pain, and the pain experienced following surgical procedures.

The primary studies included in the analysis are reported in Table 12. There were nine studies reporting a total of *N* = 730 participants. The analysis included participants from the age of three to 16 years. The study by Yeh (2005) reported correlations for multiple age groups, so each of these groups was included in the synthesis separately. The random effects model in Figure 5 was calculated using the generic inverse variance method. The random effects model suggested a weighted average correlation of $r = .65$, 95% CI [.56, .73]. This suggests a moderate positive correlation between self-report of pain and observation using the FLACC.

The level of heterogeneity in the portions reported in the primary studies was found to be within acceptable parameters for the current review (tau$^2$ = .04, Higgin's I$^2$ = 69.6%; $Q = 32.85$, $p < .001$).

*Table 12*: Methodological details and quality rating of primary studies reporting a correlation between pain scores obtained from FLACC and self-report. Details presented as reported in published paper.

| **Study** | ***n*** | **Age range (years)** | **Sample** | **Pain** | **Blinding** | **Observation** | **Self-report** | **Source of Pain** | **CHEOPS rater(s)** | **Self-report scale** | ***r*** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Berberich & Landman (2009) | 41 | 4-6 | 1 | 1 | 1 | 2 | 2 | Vaccination | Investigator | FPSR | .74 |
| Bjorkman et al. (2012) | 29 | 5-15 | 1 | 1 | 1 | 2 | 2 | Radiography following fracture | Researcher | Colour Analogue Scale | .63 |
| Bringuier et al (2009) | 19 | 4-7 | 1 | 1 | 3 | 2 | 2 | Surgery | Nurses/ Anesthetists | FPSR | .51 |
| Chadha et al. (2013) | 69 | 3-12 | 1 | 2 | 3 | 2 | 2 | Nasendoscopy | Observer | WBF | .63 |
| Elbay et al (2015) | 59 | 6-12 | 1 | 3 | 1 | 2 | 3 | Delivery of dental anaesthesia | Dentist | WBF | .39 |
| Emmott et al (2017) | 112 | 3-6 | 1 | 2 | 1 | 2 | 1 | Venipuncture | Observer | S-FPS | .74 |
| Nilsson et al. (2008) | 80 | 5-16 | 1 | 2 | 1 | 3 | 3 | Cannulation | Nurse | CAS | .61 |
| Risaw et al (2017) | 210 | 4-6 | 0 | 1 | 1 | 3 | 2 | Blood sampling | Researcher | WBF | .79 |
| Yeh (2005; 3-4yrs) | 32 | 3-4 | 1 | 1 | 0 | 2 | 3 | Surgery | Not named | Oucher | .59 |
| Yeh (2005; 5yrs) | 35 | 5 | 1 | 1 | 0 | 2 | 3 | Surgery | Not named | Oucher | .75 |
| Yeh (2005; 6yrs) | 44 | 6 | 1 | 1 | 0 | 2 | 3 | Surgery | Not named | Oucher | .46 |

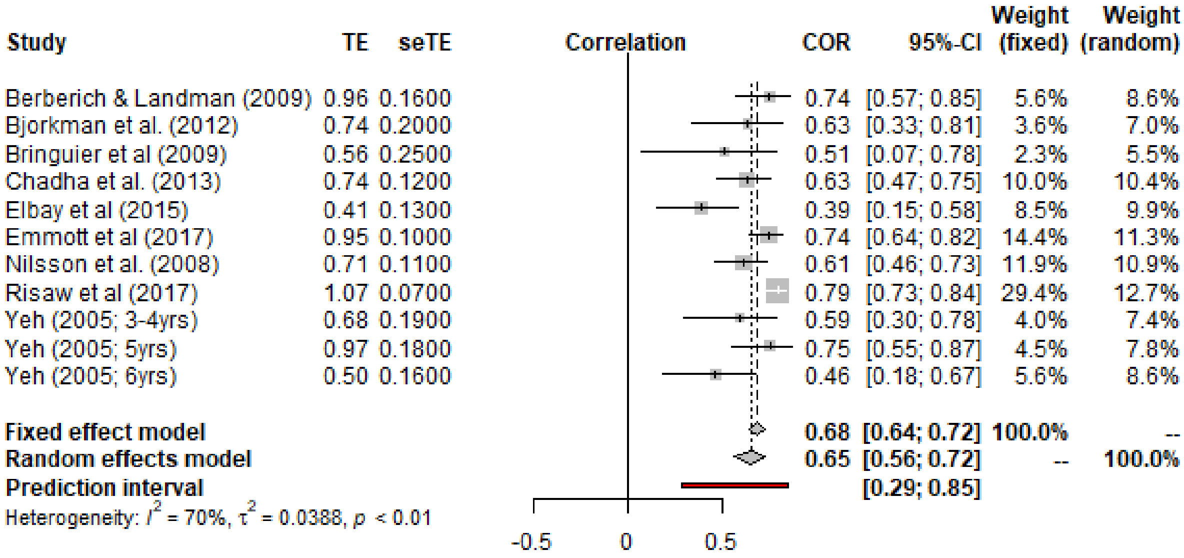


*Figure 5*: Forest plot illustrating the meta-analytic synthesis of correlations between pain ratings obtained by FLACC and self-report. TE = Measure of the effect, transformed into *z* score; seTE = measure of standard error; COR = *r* score.

The quality effect model reported a synthesis of *r* = .63, 95%CI [.53, .71]. The quality effects model evidences an approximately 3.24% decrease relative to the random effects estimate. Accordingly, when the synthesis includes information about the methodological quality of the studies there is no important change in the synthesis of these study.

As can be seen from Figure 6 the funnel plot shows asymmetry in the published studies. A trim and fill procedure was undertaken to adjust the symmetry of the funnel plot. The uncorrected estimate of the effect size is *r* = .68, the adjusted estimate is *r* = .71, 95% CI [.63, .78]. The adjusted point estimate suggests greater effect than the original analysis. The Rosenthal algorithm suggests a failsafe number of 1532 unpublished null effect studies required to reduce the meta-analytic effect of the nine results reported here.

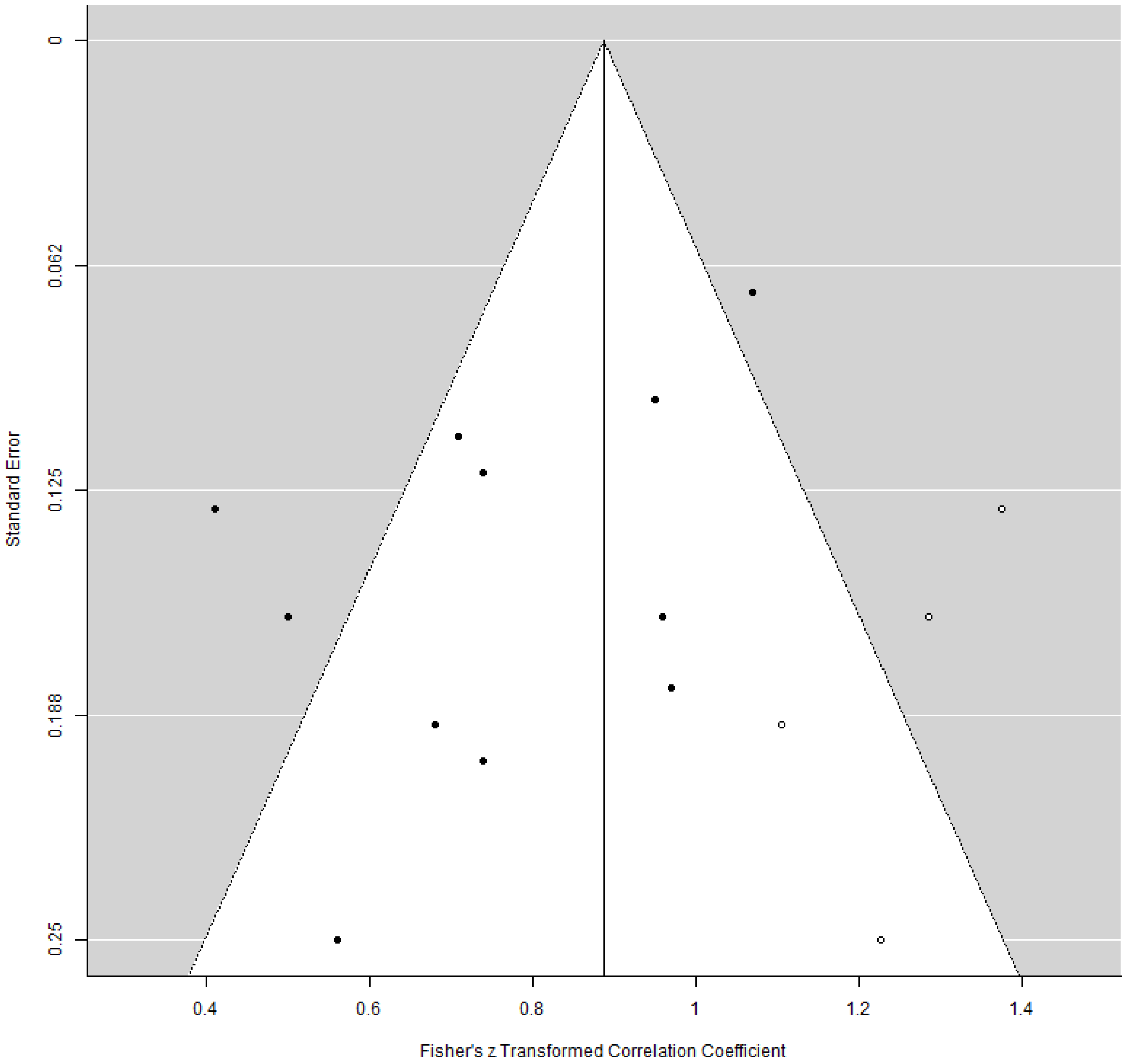


*Figure 6*: Funnel plot illustrating trim and fill procedure for FLACC analysis. Black dots indicate primary studies, white dots indicate studies imputed by trim and fill procedure.

Subgroup analyses by quality rating, presented in Table 13, suggest a significant effect of sample selection, with slightly higher correlations reported from studies categorised as “unable to rate” than those categorised as “low quality” on the sample selection domain. However the effect of sample selection may be discounted based on the disproportionate spread of studies between groups. The additional subgroup analyses presented in Table 14 found no significant results based on the role of the person completing the FLACC, or type of pain being rated.

*Table 13*: Subgroup analyses of primary studies reporting FLACC grouped by quality rating for each quality domain.

| | **Not able to rate (*k*)** | **Low quality (*k*)** | **Moderate quality (*k*)** | **High quality (*k*)** | ***Q*** | ***p*** |
|---|---|---|---|---|---|---|
| Sample selection | .79<br>(1) | .62<br>(10) | -<br>(0) | -<br>(0) | 12.53 | < .001 |
| Cause of pain | -<br>(0) | .67<br>(7) | .67<br>(3) | .39<br>(1) | 7.78 | .020 |
| Use of blinding | .61<br>(3) | .67<br>(6) | -<br>(0) | .61<br>(2) | 0.67 | .717 |
| Use of observer measure | -<br>(0) | -<br>(0) | .63<br>(9) | .72<br>(2) | 0.74 | .391 |
| Use of self-report | -<br>(0) | .74<br>(1) | .70<br>(5) | .56<br>(5) | 5.69 | .058 |

*Table 14*: Subgroup analyses of primary studies reporting FLACC grouped by type of pain rated and role of person rating FLACC.

| **Type of Pain** | **Acute Procedural (*k* = 6)** | **Post – Surgical (*k* = 4)** | **Other (*k* = 1)** | ***Q*** | ***p*** |
|---|---|---|---|---|---|
| | .67 | .59 | .63 | 0.80 | .671 |
| **FLACC rater** | **Clinician (*k* = 3)** | **Researcher (*k* = 5)** | **Not named (*k* = 3)** | | |
| | .52 | .73 | .61 | 8.44 | .015 |

#### 1.3.4.4. Observer Pain Scale.

The Observer Pain Scale (OPS; Hannallah et al., 1987) scale was found reported under a range of names, including the Hannallah Pain Scale, the Broadman Pain Scale, and the Observer Pain and Distress Scale. The measure was originally reported in a 1987 study comparing the effectiveness of different nerve block techniques (Hannallah et al., 1987). However, an evaluation of the psychometrics of the OPS was not published until a year later (Broadman, Rice & Hannallah, 1988), hence both the 1987 and 1988 references are found reported in the literature, though both refer to the same scale. The original scale rates blood pressure and four observed behaviours, each on a 0 - 2 scale, to produce a total score of 0 - 10. Some studies choose to omit the blood pressure measurement and rely only on the four behaviours with a 0 - 8 total scale.

Only one study was available which reported correlations between the OPS and self-report of pain. Bringuier et al (2009) is also described in the analysis for the CHEOPS and the FLACC. The study presents an investigation into the efficacy of behavioural pain tools and so utilized four different behavioural assessments alongside the FPSR as a self-report measure. The study reported using the OPS without the item relating to blood pressure. Although the original sample reported by Bringuier et al. (2009) is *N* = 150, the correlation reported for the time point which is closest to the painful procedure is based on *n* = 19. The reported correlation between the OPS and the FPSR is *r* = .64.

#### 1.3.4.5. Wong Baker Faces.

Although multiple versions of the faces pain scale were reported, the Wong-Baker Faces Scale (WBF; Wong & Baker, 1988) was the most frequently used. The WBF was originally published as a self-report scale for children. A series of six cartoon faces ranging from smiling to crying are depicted with verbal anchors ranging from “No pain” to “Hurts worst”. Each face corresponds to a numerical score, increasing in twos, from 0 - 10. The majority of studies included in this review that used a WBF reported it as a self-report scale as well as obtaining ratings on the WBF from observers.

The primary studies included in the analysis are reported in Table 15. There were four studies reporting a total of *N* = 369 participants. The analysis included participants from the age of three to  15 years.

The random effects model in Figure 7 was calculated using the generic inverse variance method. The random effects model suggested a weighted average correlation of *r* = .86 and a 95% CI [.51, .97]. This suggests a moderate positive correlation between self-report of pain and observation using the Wong-Baker Faces.

*Table 15*: Methodological details and quality rating of primary studies reporting a correlation between pain scores obtained from WBF and self-report. Details presented as reported in published paper.

| **Study** | ***n*** | **Age range (years)** | **Sample** | **Pain** | **Blinding** | **Observation** | **Self-report** | **Source of Pain** | **CHEOPS rater(s)** | **Self-report scale** | ***r*** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Chadha et al. (2013) | 69 | 3-12 | 1 | 1 | 1 | 1 | 2 | Nasendoscopy | Parent | WBF | .68 |
| Moadad et al. (2015) | 48 | 4-12 | 2 | 1 | 1 | 1 | 3 | IV insertion | Nurse | WBF | .37 |
| Mohan et al.  (2015) | 42 | 10-15 | 1 | 1 | 1 | 1 | 2 | "painful procedures" | Nurse | WBF | .98 |
| Risaw et al (2017) | 210 | 4-6 | 0 | 1 | 1 | 1 | 2 | Blood sampling | Researcher | WBF | .94 |

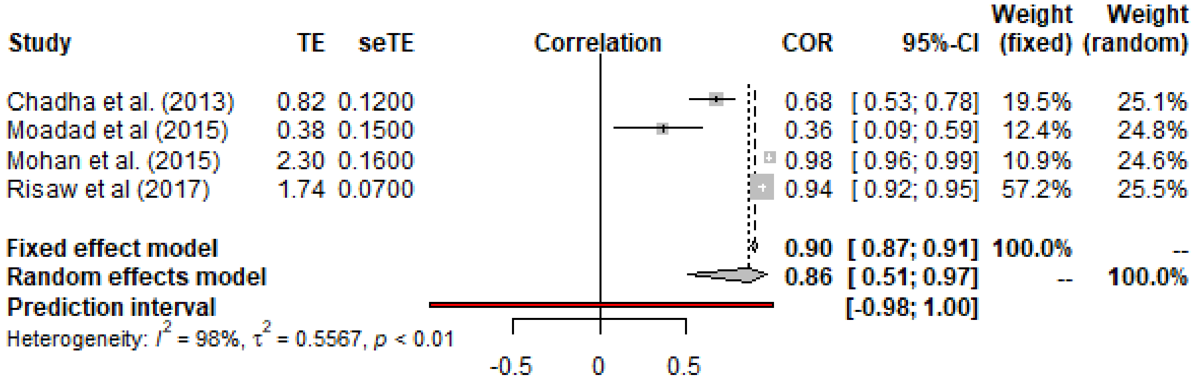


*Figure 7*: Forest plot illustrating the meta-analytic synthesis of correlations between pain ratings obtained by WBF and self-report. TE = Measure of the effect, transformed into Fisher's z score; seTE = measure of standard error; COR = *r* score.

An unacceptable level of heterogeneity in the portions reported in the primary studies was observed (tau$^2$ = .56, Higgin's I$^2$ = 97.6%; Q = 123.83, $p$ < 0.0001). This suggests that the estimates of the primary studies are biased by the presence of uncontrolled or confounding factors. The quality effect model reported a synthesis of *r* = 0.84, 95% CI [.43, .96]. The quality effects model evidences an approximately 2.92% decrease relative to the random effects estimate.

Despite excessive levels of heterogeneity being identified, none of the studies met the criterion for removal therefore no corrections were made to the analysis based on the leave one out analysis. Because of the high levels of heterogeneity identified and the small number of primary studies available for the synthesis of the WBF, the funnel plot and sub-group analyses were not conducted as results would not have been meaningful.

### 1.3.4.6. Comparison between observational measures.

The results of the conducted meta-analyses are summarised in Table 16. Moderate to strong associations with self-report scores were identified for all observational assessments. Unacceptable levels of heterogeneity were identified in two of the analyses (VAS, $I^2 = 85.8\%$; WBF, $I^2 = 97.6\%$). Highest correlations between observer ratings and self-report were identified for the WBF scale.

*Table 16*: Results of meta-analytic syntheses of correlations between pain scores obtained from observer tools and those obtained from self-report, compared across the five most frequently used observer tools found in the literature review.

| | | | | | **Heterogeneity** | | | |
|---|---|---|---|---|---|---|---|---|
| **Measure** | ***K*** | **Total *N*** | **Age range (years)** | **Weighted *r* value [95% CI]** | **Tau$^2$** | **I$^2$** | ***Q*** | ***p*** |
| CHEOPS | 8 | 517 | 4-15 | .52<br>[.39, .64] | .04 | 65.2% | 23.00 | .003 |
| VAS | 14 | 1188 | 3-18 | .62<br>[.51, .70] | .07 | 84.4% | 83.54 | < .001 |
| FLACC | 9 | 730 | 3-16 | .65<br>[.55, 0.73] | .04 | 69.6% | 32.85 | < .001 |
| OPS | 1 | 19 | 4-7 | .64 | - | - | - | - |
| WBF | 4 | 369 | 3-15 | .86<br>[.51, .97] | .56 | 97.6% | 123.83 | < .001 |

## 1.5. Discussion

The current review aimed to identify the observational tools utilised in the published literature to assess pain in children. A total of 62 unique observational assessments were found reported across 526 papers published from 1979-2018. The second aim of the current study was to evaluate the evidence of validity for commonly used measures of pain by synthesising the available data regarding correlation to self-report, the current gold-standard in pain assessment. The five most commonly reported observational assessments of pain were found to have moderate to strong positive correlations to self-report, though the availability of these data varied considerably between measures. The current review is the first to apply meta-analytic methods to assess the validity of current pain assessment methods in research using child participants. This meta-analysis is strengthened by the comprehensive search, and assessment of the five most commonly used measures. Given that nearly two thirds of the identified literature reported using at least one of the five most common measures (65.78%), the current review has far reaching implications for research and practice.

A total of 62 tools were identified across the 526 papers reviewed; Anderson et al. (2017) identified 65 measures using broadly similar inclusion criteria for their literature review. The slight difference in the number of identified measures may be due to differences in the categorisation of techniques such as global rating scales, or measures that were designed for the needs of one specific study. Both Anderson et al., and Von Baeyer and Spagrud (2007) highlight a lack of published evidence regarding the psychometric properties of the observational measures used to assess pain in children. The sheer number of measures identified is a concern in this regard; such a range of measures introduces heterogeneity and confusion to the literature and makes it difficult to compare findings across studies. This is particularly the case for the 29 measures that were reported at very low frequency (*k* = 1) in the current review. The limited use of some of the reported measures is easily explained because of the exclusion criteria of the current review, for example the Comfort scale (Ambuel, Hamlett, Marx & Blumer, 1992) was reported in only four studies, but is designed for use in very young infants, and is reported much more frequently in studies assessing pain in infants under 12 months (Duhn & Medves, 2004). Similarly, the Echelle Douleur Enfant San Salvador (Collignon & Giusiano, 2001), was reported in only one study, but is designed to assess pain in children with cerebral palsy, a group which was excluded from the current review. However, this explanation does not apply to all of the measures reported at low frequency, and 13 papers were identified in which authors created their own measure for the purposes of the study, rather than using already established and validated measures. In these cases it is difficult to justify the use measures that have little record of publication and therefore lack robust evidence of validity or reliability. Unless there is clear justification for why published measures fail to meet the needs of the study, researchers should avoid adding to this already crowded picture.

The second aim of the review was to evaluate the convergence between self-report and the scores obtained from the five most commonly used observational measures of pain. The five measures identified as being most commonly reported in the literature were the Children's Hospital East Ontario Pain Scale (CHEOPS; McGrath et al., 1985) which was used in 121 studies, the Visual Analogue Scale

(VAS) which was used in 87 studies, the Face, Legs, Activity, Crying and Consolability (FLACC; Voepel-Lewis, Shayevitz & Malviya, 1997) which was used in 81 studies, the Observer Pain Scale (OPS; Hannallah et al., 1987) which was used in 71 studies, and the Wong-Baker Faces Scale (WBF; Wong & Baker, 1988) which was used in 26 studies. Based on the categories of measures provided by Von Baeyer & Spagrud (2007), the CHEOPS, FLACC, and OPS, fit the description of behavioural rating scales or behavioural checklists, whereas the VAS and WBF are both considered global rating scales.

### 1.5.1. Children's Hospital East Ontario Pain Scale

The CHEOPS was the most commonly reported observational measure and the synthesis of correlations revealed a moderate positive correlation with self-report measures across the eight available primary studies (*r* = .52). Subgroup analyses revealed an effect of rater, suggesting that CHEOPS ratings were more closely associated with self-report of pain when the observer using the CHEOPS was a researcher rather than a clinician. It may be that researchers are more likely to comply rigidly to the definitions given by a measure, whereas clinicians may have more of a tendency to alter their ratings based upon clinical experience. This finding must be treated with caution however, not only because of the small number of studies included in the analysis, but also because of the lack of clarity between the categories of rater. Raters were classified as "researcher" when the study described them as researchers or observers as opposed to using a clinical job title, however this does not exclude the possibility that raters included in the researcher category may also have been clinically trained. If further data support the finding that researchers provide CHEOPS ratings which are closer to self-report than ratings provided by clinicians, then studies seeking to use the CHEOPS as a proxy for self-report may be better placed to use researchers to provide such ratings, or to give explicit instruction or training to others using the measure.

A review by the Pediatric Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (PedIMMPACT) group regarding the use of observational measures of pain recommends the use of the CHEOPS for clinical trials, but only for the assessment of acute pain (Von Baeyer & Spagrud, 2007). The current review would support the recommendations of the PedIMMPACT group and suggests that published research is broadly in line with best practice, in that the CHEOPS was identified as the most commonly used measure. It should be noted however that contrary to PedIMMPACT recommendations, almost half of the primary studies in the CHEOPS synthesis used the measure to asses post-surgical pain, practice which is not supported by PedIMMPACT recommendations. The current review attempted to assess if the association between observed pain scores on the CHEOPS and self-reported pain was poorer in studies assessing post-surgical pain, using sub-group analysis methods. No significant difference was found in the correlation with self-report between those studies assessing procedural pain and those assessing post-surgical pain, which does not appear to support the caution expressed by PedIMMPACT. It is also of note that although the weighted correlation calculated between the CHEOPS and self-report was of moderate strength, it was the weakest of the five measures assessed.

Although it is promising that the CHEOPS was the most frequently identified measure, and that the meta-analysis suggests a moderate positive correlation to self-report, further data are needed to assess whether the validity and reliability of the CHEOPS is still acceptable when the measure is applied to post-surgical pain.

### 1.5.2. Visual Analogue Scale

The VAS was the second most commonly reported observational measure. Unlike the CHEOPS, the VAS is a global rating scale, which does not use specific behavioural indicators to guide or justify ratings. The VAS was found to have a moderate positive correlation to self-report (*r* = .62), however this was only slightly stronger than the correlation calculated for the CHEOPS, which was the weakest correlation of the five. One strength of the VAS is that it can also be used as a self-report scale; there is evidence that children as young as 6 years can reliably self-report using the VAS (Von Baeyer, 2006). It was thought that the use of the same scale for self-report and observer report might be one factor explaining the correlation between the VAS and self-report. A sub-group analysis found that there was no difference in correlation found between studies that used the VAS for self-report and those that used a different self-report scale. However, it is also arguable that almost all of the self-report scales used in the primary studies reported here are simply variations of a VAS, the differences being in the choice of anchor points and the use of visual aids, such as pictures of faces, to aid rating choice.

Unlike the behavioural rating scales, no effect of rater was found for the VAS. The sub group analysis of rater for the VAS included clinicians, researchers, and parents assessing pain in their own children, suggesting the VAS performs consistently regardless of the clinical knowledge of the person using it, or their knowledge of the child being rated. Although overall observer ratings obtained from the VAS correlated well with self-report, there was a wide range of correlations reported across the primary studies included (*r* = .18 - .82), and unacceptably high levels of heterogeneity between studies. The high levels of heterogeneity identified in the primary studies limit the conclusions that might be drawn from this meta-analysis. High levels of heterogeneity suggest that the findings of the studies cannot be reliably attributed to idiosyncratic variation in the correlation, but are more likely related to methodological factors. The leave one out analysis failed to identify a single study that made a significant contribution to the heterogeneity, and sub-group analyses of study quality also failed to return significant findings. With no clear source of heterogeneity identified in the current review, it is difficult to draw any robust conclusions from the synthesis of studies using the VAS.

### 1.5.3. Face Legs Activity Crying Consolability

The FLACC was identified as the third most frequently used measure in the current review. The FLACC was found to have a moderate-strong positive correlation to ratings of pain obtained by self-report. The weighted correlation found for the FLACC was higher than that found for either of the other behavioural scales (the CHEOPS and OPS) explored in the current review, though it performed slightly poorer than the WBF.

Unlike the CHEOPS, PedIMMPACT recommends the use of the FLACC for the assessment of both acute procedural and post-surgical pain in clinical trials. The FLACC also contains fewer items and appears to be a simpler tool to administer than the CHEOPS. Given that the FLACC can be applied to a broader range of settings and appears to perform better than the CHEOPS regarding association with self-report, it begs the question why the CHEOPS is used more frequently. The finding that the CHEOPS is used more frequently than the FLACC may be explained by the breadth of the current review with regards to year of publication. The CHEOPS was first published in 1985, a full 12 years before the FLACC. This is reflected in the publication year of the primary studies, the primary studies included in the meta-analysis of the CHEOPS were all published between 1992 and 2011. In contrast, the primary studies included in the meta-analysis of the FLACC were more recent, all having been published since 2005. It may be that modern researchers are indeed showing a preference for the FLACC but this has not yet been sufficient to overtake the CHEOPS due to the historical primacy of the CHEOPS. The results of this meta-analysis would support the use of the FLACC where an observational assessment of pain is required, given the positive correlation between the FLACC and self-report, the simplicity of the measure, and the variety of settings for which it has been validated.

### 1.5.4. Observer Pain Scale

Despite being widely used in the reviewed literature, the OPS was rarely compared to self-report. Only 11.27% of studies using the OPS included a measure of self-report ($k$ = 8), and in only one of those was an analysis of the association between the two measures reported.

The first aim of the current review was to identify the observational measures used to assess pain in children in the literature, however this seemingly simple task was complicated in some cases by multiple measures being referred to by similar names, or by a single measure being referred to by multiple names. The OPS was one of the most prolific examples of this. The OPS was found referred to as the Observer Pain Scale, the Hannallah Pain Scale, the Broadman Pain Scale, and the Observer Pain and Distress Scale. There were also variations found in scoring and administration, most commonly the omission of the item regarding blood pressure, however the scoring variations appeared to bear no relation to the different names. Inconsistent reporting of measures in research increases the challenges when attempting to synthesise the literature.

Although a meta-analysis of the literature could not be conducted for the OPS, the correlation between the OPS and self-report found by Bringuier et al. (2009) was in line with the correlations found between other scales and self-report (*r* = .52 - .86). The findings of the current review suggest that the OPS compares well with other observational measures of pain in relation to its correlation to self-report, however the correlation is based on a very small sample and so must be interpreted with caution.

### 1.5.5. Wong Baker Faces

The WBF was the least used of the five measures reviewed. The synthesis of the four available studies meeting inclusion criteria suggest that the WBF, when used as an observer scale, achieves the highest correlation to self-report (*r* = .86) across the five tools reviewed. However, the synthesis of the WBF was also the synthesis with the most heterogeneity. Similar to the VAS, the four studies included reported a wide range of correlations, and no clear sources of heterogeneity could be identified. Because of the small number of primary studies, sub-group analyses were not conducted. All of the primary studies reporting the WBF as an observational measure also used the WBF for collection of self-report. Because no studies were identified that used the WBF as an observational measure with a different self-report tool, the effect of using the same tool for observation and self-report could not be assessed in the case of the WBF. Despite finding a strong positive correlation to self-report, the current review concludes that the current literature regarding the WBF is too variable to support a recommendation for the use of the measure as a proxy for self-report in clinical or research practice.

## 1.6. Limitations

The current review extends current understanding of the validity of commonly used observational assessments of pain by exploring the association between scores obtained from observation and those obtained by self-report, which is considered the gold-standard. Correlation to self-report offers some insight to validity, however there are limitations to the degree of variability that may be detected using correlation analysis, and the approach is reliant on the assumption that self-report of pain is an accurate measure of pain intensity. Despite being considered the gold-standard, self-report as a measure of pain has limitations, particularly when applied to children. Firstly one must be careful to ensure that children can understand and engage in the self-report tool used. In the current review, steps were taken to ensure the validity of self-report, for example, by the exclusion of studies with children under three years of age, who are unlikely to be reliable in their reporting of pain. Use of self-report, in particular the application of a self-report tool that had been validated for use in the named sample, was also a domain in the quality rating criteria. However, many of the primary studies received low ratings in this quality domain. Even if self-report provided a perfect measure of pain intensity, correlation provides only limited insight into the differences in intensity ratings between observation and self-report. For example, even where significant correlations have been reported between child and parent scores on the VAS, agreement between pain ratings has been found to be poor (Kelly, Powell & Williams, 2002). So long as children are ranked in appropriate order with regard to which children appear to experience the most pain, a correlation analysis will not detect disagreements in the rating of pain intensity between the child and observer.

Correlation is not the only available method to assess relationships between two measures, however it was the most widely reported. Some of the papers that utilised observation measures alongside self-report did conduct comparisons by defining thresholds for scores to group participants into those experiencing no pain, or moderate-severe pain. However, the variations in methodology and reporting between such studies would have made a meta-analytic synthesis of the data very difficult. Correlation became the clear option for the current review because of the availability of data and because of the relative consistency in the reporting of correlations. Unfortunately some papers were excluded from the meta-analysis because, although they reported conducting a correlation analysis, they failed to report an r value. Some studies without an r value were those reporting no significant correlation between observer and self-report, which may suggest publication bias within the literature. However, it was also the case that r values had to be calculated for two studies who reported a *p* value without reporting an r value (Hee et al., 2003;

Benini et al., 2004). In both cases the calculated r value was considerably lower than other r values included in the synthesis, this may be a product of the transformation calculation; however it may also suggest why the authors chose to omit the r value in their reporting, instead opting to simply report the value as being significant. The synthesis reported here was the most comprehensive possible without conducting significant transformations of the published data. Future studies must ensure that data is reported in full, including non-significant findings, although it will not negate the problem of publication bias, it will allow for future reviews to be more representative of the full depth of knowledge available on these measures which are potentially so important in the development of better medical techniques.

## 1.7. Conclusions

The current review was conducted to explore the use of observational assessments to assess pain in children in research literature. A total of 62 measures were identified, suggesting a great deal of inconsistency in current practice regarding the assessment of pain in children participating in clinical studies.

Two of the five most frequently used observational measures, the VAS and WBF, can be categorised as global rating scales. The results reported here suggest that global rating scales may have a higher correlation to self-report but also present unacceptably high levels of heterogeneity, limiting the usefulness of the analysis. It may be that heterogeneity is unavoidable when using global rating scales due to the lack of guidance regarding scoring criteria, potentially increasing the influence of individual factors such as user experience or bias. The FLACC, a behavioural checklist, was found to have a weighted correlation higher than the VAS, and heterogeneity in the literature was found to be lower than in the case of either the VAS or WBF.

The results of the current study support the use of observational scales by providing evidence of a positive association to self-report, however the variability found within the literature was of concern. Recommendations for future research would be to restrict the use of observational measures of pain to those measures that are already well established in the literature, and to provide further exploration and evidence of the use of global rating scales as observer measures, as this study demonstrates that such measures are currently in frequent use despite a lack of such evidence.

## 1.8. References

Ambuel, B., Hamlett, K. W., Marx, C. M., & Blumer, J. L. (1992). Assessing distress in pediatric intensive care environments: the COMFORT scale. *Journal of pediatric psychology*, *17*(1), 95-109.

Anand, K. J., & Craig, K. D. (1996). New perspectives on the definition of pain. *Pain-Journal of the International Association for the Study of Pain*, *67*(1), 3-6.

Andersen, R. D., Langius-Eklöf, A., Nakstad, B., Bernklev, T., & Jylli, L. (2017). The measurement properties of pediatric observational pain scales: A systematic review of reviews. *International journal of nursing studies*, *73*, 93-101.

Attia, J., Amiel-Tison, C., Mayer, M. N., Shnider, S. M., & Barrier, G. (1987). Measurement of postoperative pain and narcotic administration in infants using a new clinical scoring system. Anesthesiology: *The Journal of the American Society of Anesthesiologists, 67*(3), A532-A532.

Beadle-Brown, J., Murphy, G., & DiTerlizzi, M. (2009). Quality of life for the Camberwell cohort. *Journal of Applied Research in Intellectual Disabilities*, *22*(4), 380-390.

Bearden, D. J., Feinstein, A., & Cohen, L. L. (2012). The influence of parent preprocedural anxiety on child procedural pain: mediation by child procedural anxiety. *Journal of pediatric psychology*, *37*(6), 680-686.

Benini, F., Trapanotto, M., Gobber, D., Agosto, C., Carli, G., Drigo, P., ... & Zacchello, F. (2004). Evaluating pain induced by venipuncture in pediatric patients with developmental delay. *The Clinical journal of pain*, *20*(3), 156-163.

Berberich, F. R., & Landman, Z. (2009). Reducing immunization discomfort in 4-to 6-year-old children: a randomized clinical trial. *Pediatrics*, *124*(2), e203-e209.

Beyer, J. E., McGrath, P. J., & Berde, C. B. (1990). Discordance between self-report and behavioral pain measures in children aged 3–7 years after surgery. *Journal of Pain and Symptom Management*, *5*(6), 350-356.

Beyer, J. E., Denyes, M. J., &  Villarruel, A., M. (1992) "The creation, validation, and continuing development of the Oucher: a measure of pain intensity in children." *Journal of pediatric nursing* 7.5, 335-346.

Bieri, D., Reeve, R. A., Champion, G. D., Addicoat, L., & Ziegler, J. B. (1990). The Faces Pain Scale for the self-assessment of the severity of pain experienced by children: development, initial validation, and preliminary investigation for ratio scale properties. *Pain*, *41*(2), 139-150.

Björkman, B., Nilsson, S., Sigstedt, B., & Enskär, K. (2012). Children's pain and distress while undergoing an acute radiographic examination. *Radiography*, *18*(3), 191-196.

Boivin, J. M., Poupon-Lemarquis, L., Iraqi, W., Fay, R., Schmitt, C., & Rossignol, P. (2008). A multifactorial strategy of pain management is associated with less pain in scheduled vaccination of children. A study realized by family practitioners in 239 children aged 4–12 years old. *Family practice*, *25*(6), 423-429.

Breau, L. M., McGrath, P. J., Craig, K. D., Santor, D., Cassidy, K. L., & Reid, G. J. (2001). Facial expression of children receiving immunizations: a principal components analysis of the child facial coding system. *The Clinical journal of pain*, *17*(2), 178-186.

Bringuier, S., Picot, M. C., Dadure, C., Rochette, A., Raux, O., Boulhais, M., & Capdevila, X. (2009). A prospective comparison of post-surgical behavioral pain scales in preschoolers highlighting the risk of false evaluations. *PAIN*, *145*(1-2), 60-68.

Broadman, L. M., Rice, L. J., & Hannallah, R. S. (1988). Evaluation of an objective pain scale for infants and children. *Regional Anesthesia and Pain Medicine*, *13*(1), 45.

Büttner, W., Breitkopf, L., Miele, B., & Finke, W. (1990). Initial results of the reliability and validity of a German-language scale for the quantitative measurement of postoperative pain in young children. *Der Anaesthesist, 39*(11), 593-602.

Büttner, W., & Finke, W. (2000). Analysis of behavioural and physiological parameters for the assessment of postoperative analgesic demand in newborns, infants and young children: a comprehensive report on seven consecutive studies. *Pediatric Anesthesia, 10*(3), 303-318.

Cameron, E., Johnston, G., Crofts, S., & Morton, N. S. (1992). The minimum effective dose of lignocaine to prevent injection pain due to propofol in children. *Anaesthesia, 47*(7), 604-606.

Carr, E. G., & Owen-DeSchryver, J. S. (2007). Physical illness, pain, and problem behavior in minimally verbal people with developmental disabilities. *Journal of Autism and Developmental Disorders*, *37*(3), 413-424.

Cassidy, K. L., Reid, G. J., McGrath, P. J., Finley, G. A., Smith, D. J., Morley, C., ... & Morton, B. (2002). Watch needle, watch TV: Audiovisual distraction in preschool immunization. *Pain Medicine*, *3*(2), 108-118.

Chadha, N. K., Lam, G. O., Ludemann, J. P., & Kozak, F. K. (2013). Intranasal topical local anesthetic and decongestant for flexible nasendoscopy in children: a randomized, double-blind, placebo-controlled trial. *JAMA Otolaryngology–Head & Neck Surgery*, *139*(12), 1301-1305.

Chambers, C. T., Reid, G. J., McGrath, P. J., & Finley, G. A. (1996). Development and preliminary validation of a postoperative pain measure for parents. *Pain, 68*(2-3), 307-313.

Cohen, L. L., Blount, R. L., Cohen, R. J., & Johnson, V. C. (2004). Dimensions of pediatric procedural distress: Children's anxiety and pain during immunizations. *Journal of Clinical Psychology in Medical Settings*, *11*(1), 41-47.

Collignon, P., & Giusiano, B. (2001). Validation of a pain evaluation scale for patients with severe cerebral palsy. *European Journal of Pain*, *5*(4), 433-442.

Craig, K. D. (2009). The social communication model of pain. *Canadian Psychology / Psychologie canadienne*, *50*(1), 22.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled clinical trials*, *7*(3), 177-188.

Doi, S. A., & Thalib, L. (2008). A quality-effects model for meta-analysis. *Epidemiology*, 94-100.

Duhn, L. J., & Medves, J. M. (2004). A systematic integrative review of infant pain assessment tools. *Advances in Neonatal care*, *4*(3), 126-140.

Duval, S. & Tweedie, R. (2000a), A nonparametric “Trim and Fill” method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*, 89-98.

Duval, S. & Tweedie, R. (2000b), Trim and Fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56,* 455-463.

Ekamn, P., & Friesen, W. (1978). *Facial action coding system (FACS): manual*.

Elbay, M., Şermet, Ü. E., Yıldırım, S., Uğurluel, C., Kaya, C., & Baydemir, C. (2015). Comparison of injection pain caused by the DentalVibe Injection System versus a traditional syringe for inferior alveolar nerve block anaesthesia in paediatric patients. *European journal of paediatric dentistry: official journal of European Academy of Paediatric Dentistry*, *16*(2), 123-128.

Emmott, A. S., West, N., Zhou, G., Dunsmuir, D., Montgomery, C. J., Lauder, G. R., & Von Baeyer, C. L. (2017). Validity of simplified versus standard self-report measures of pain intensity in preschool-aged children undergoing venipuncture. *The Journal of Pain*, *18*(5), 564-573.

Fanurik, D., Koh, J. L., Harrison, R. D., Conrad, T. M., & Tomerun, C. (1998). Pain assessment in children with cognitive impairment: an exploration of self-report skills. *Clinical Nursing Research*, *7*(2), 103-119.

Fearon, I., McGrath, P. J., & Achat, H. (1996). 'Booboos': the study of everyday pain among young children. *Pain, 68*(1), 55-62.

Foster, R. L., & Varni, J. W. (2002). Measuring the Quality of Children's Postoperative Pain Management: Initial Validation of the Child/Parent Total Quality Pain Management (TQPM™) Instruments. *Journal of Pain and Symptom Management*, *23*(3), 201-210.

Gauvain-Piquard, A., Rodary, C., Rezvani, A., & Lemerle, J. (1987). Pain in children aged 2–6 years: a new observational rating scale elaborated in a pediatric oncology unit—preliminary report. *Pain, 31*(2), 177-188.

Gilbert, C. A., Lilley, C. M., Craig, K. D., McGrath, P. J., Bennett, S. M., & Montgomery, C. J. (1999). Postoperative pain expression in preschool children: validation of the child facial coding system. *The Clinical journal of pain, 15*(3), 192-200.

Goodenough, B., Addicoat, L., Champion, G. D., McInerney, M., Young, B., Juniper, K., & Ziegler, J. B. (1997). Pain in 4-to 6-year-old children receiving intramuscular injections: a comparison of the Faces Pain Scale with other self-report and behavioral measures. *The Clinical journal of pain, 13*(1), 60-73.

Goodenough, B., Thomas, W., Champion, G. D., Perrott, D., Taplin, J. E., von Baeyer, C. L., & Ziegler, J. B. (1999). Unravelling age effects and sex differences in needle pain: ratings of sensory intensity and unpleasantness of venipuncture pain by children and their parents. *Pain*, *80*(1-2), 179-190.

Grunau, R. V., Johnston, C. C., & Craig, K. D. (1990). Neonatal facial and cry responses to invasive and non-invasive procedures. *Pain, 42*(3), 295-305.

Güleç, S., Büyükkidan, B., Oral, N., Özcan, N., & Tanriverdi, B. (1998). Comparison of caudal bupivacaine, bupivacaine-morphine and bupivacaine-midazolam mixtures for post-operative analgesia in children. *European journal of anaesthesiology, 15*(2), 161-165.

Hannallah, R. S., Broadman, L. M., Belman, A. B., Abramowitz, M. D., & Epstein, B. S. (1987). Comparison of caudal and ilioinguinal/iliohypogastric nerve blocks for control of post-orchiopexy pain in pediatric ambulatory surgery. *Anesthesiology: The Journal of the American Society of Anesthesiologists*, *66*(6), 832-833.

Hee, H. I., Goy, R. W. L., & Ng, A. S. B. (2003). Effective reduction of anxiety and pain during venous cannulation in children: a comparison of analgesic efficacy conferred by nitrous oxide, EMLA and combination. *Pediatric Anesthesia*, *13*(3), 210-216.

Hesselgard, K., Larsson, S., Romner, B., Strömblad, L. G., & Reinstrup, P. (2007). Validity and reliability of the behavioural observational pain scale for postoperative pain measurement in children 1–7 years of age. *Pediatric Critical Care Medicine, 8*(2), 102-108.

Hester, N. O. (1979). Measurements of pain in children: generalizability and validity of the pain ladder and the poker chip tool. *Pediatric pain*, 79-94.

Hicks, C. L., von Baeyer, C. L., Spafford, P. A., van Korlaar, I., & Goodenough, B. (2001). The Faces Pain Scale–Revised: toward a common metric in pediatric pain measurement. *Pain*, *93*(2), 173-183.

Hunt, A., Goldman, A., Seers, K., Crichton, N., Mastroyannopoulou, K., Moffat, V., Brady, M., et al. (2004). Clinical validation of the paediatric pain profile. Developmental medicine and child neurology, 46(1), 9-18.

Jensen, B. (2012). Post-operative pain and pain management in children after dental extraction under general anaesthesia. *European Archives of Paediatric Dentistry*, *13*(3), 119-125

Jongudomkarn, D., Angsupakorn, N., & Siripul, P. (2008). The development and validation of the Khon Kaen University Pediatric Pain Assessment Tool for school-aged Isaan children in Thailand. *Journal of Transcultural Nursing, 19*(3), 213-222.

Joyce, B. A., Schade, J. G., Keck, J. F., Gerkensmeyer, J., Raftery, T., Moser, S., & Huster, G. (1994). Reliability and validity of preverbal pain assessment tools. *Issues in Comprehensive Pediatric Nursing, 17*(3), 121-135.

Jylli, L., & Olsson, G. L. (1995). Procedural pain in a paediatric surgical emergency unit. *Acta paediatrica*, *84*(12), 1403-1408.

Kelly, A. M., Powell, C. V., & Williams, A. (2002). Parent visual analogue scale ratings of children's pain do not reliably reflect pain reported by child. *Pediatric emergency care*, *18*(3), 159-162.

Knutsson, J., Tibbelin, A., & Von Unge, M. (2006). Postoperative pain after paediatric adenoidectomy and differences between the pain scores made by the recovery room staff, the parent and the child. *Acta oto-laryngologica*, *126*(10), 1079-1083.

Krane, E. J., Jacobson, L. E., Lynn, A. M., Parrot, C., & Tyler, D. C. (1987). Caudal morphine for postoperative analgesia in children: a comparison with caudal bupivacaine and intravenous morphine. *Anesthesia and analgesia, 66*(7), 647-653.

Kuttner, L., & LePage, T. (1989). Face scales for the assessment of pediatric pain: A critical review. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement, 21*(2), 198.

Lamontagne, L. L., Johnson, B. D., & Hepworth, J. T. (1991). Children's ratings of postoperative pain compared to ratings by nurses and physicians. *Issues in Comprehensive Pediatric Nursing*, *14*(4), 241-247.

LeBaron, S., & Zeltzer, L. (1984). Assessment of acute pain and anxiety in children and adolescents by self-reports, observer reports, and a behavior checklist. *Journal of consulting and clinical psychology, 52*(5), 729.

Lee, L. W., & White-Traut, R. C. (1996). The role of temperament in pediatric pain response. *Issues in Comprehensive Pediatric Nursing*, *19*(1), 49-63.

Loeser, J. D., & Melzack, R. (1999). Pain: an overview. *The Lancet*, *353*(9164), 1607-1609.

Maunuksela, E. L., Olkkola, K. T., & Korpela, R. (1987). Measurement of pain in children with self‐reporting and behavioral assessment. *Clinical Pharmacology & Therapeutics, 42*(2), 137-141.

McGrath, P. J., Johnson, G., Goodman, J. T., Schillinger,  J., Dunn, J. & Chapman, J. (1985) CHEOPS: A behavioral scale for rating postoperative pain in children. *Advances in pain research and therapy,  9.* 395–402.McGrath, P. A., Seifert, C. E., Speechley, K. N., Booth, J. C., Stitt, L., & Gibson, M. C. (1996). A new analogue scale for assessing children's pain: an initial validation study. *Pain, 64*(3), 435-443.

Merskey, H. (1991). The definition of pain. *European psychiatry*.

Moadad, N., Kozman, K., Shahine, R., Ohanian, S., & Badr, L. K. (2016). Distraction using the BUZZY for children during an IV insertion. *Journal of pediatric nursing*, *31*(1), 64-72.

Mohan, S., Nayak, R., Thomas, R. J., & Ravindran, V. (2015). The effect of Entonox, play therapy and a combination on pain relief in children: a randomized controlled trial. *Pain Management Nursing*, *16*(6), 938-943.

Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

Nilsson, S., Finnström, B., & Kokinsky, E. (2008). The FLACC behavioral scale for procedural pain assessment in children aged 5–16 years. *Pediatric Anesthesia*, *18*(8), 767-774.

Oster, H., & Rosenstein, D. (1993). *Baby FACS: Analyzing facial movement in infants*. Unpublished manuscript, New York University.

Payen, J. F., Bru, O., Bosson, J. L., Lagrasta, A., Novel, E., Deschaux, I., Jacquot, C. et al. (2001). Assessing pain in critically ill sedated patients by using a behavioral pain scale*. Critical care medicine, 29*(12), 2258-2263.

Peters, M. L., Patijn, J., & Lamé, I. (2007). Pain assessment in younger and older pain patients: psychometric properties and patient preference of five commonly used measures of pain intensity. *Pain Medicine, 8*(7), 601-610.

Price, D. D., McGrath, P. A., Rafii, A., & Buckingham, B. (1983). The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. *Pain, 17*(1), 45-56.

Ramelet, A. S., Rees, N. W., McDonald, S., Bulsara, M. K., & Huijer Abu-Saad, H. (2007). Clinical validation of the multidimensional assessment of pain scale. *Pediatric Anesthesia, 17*(12), 1156-1165.

Risaw, L., Narang, K., Thakur, J. S., Ghai, S., Kaur, S., & Bharti, B. (2017). Efficacy of Flippits to Reduce Pain in Children during Venipuncture–A Randomized Controlled Trial. *The Indian Journal of Pediatrics*, *84*(8), 597-600.

Riva, P., Wirth, J. H., & Williams, K. D. (2011). The consequences of pain: The social and physical pain overlap on psychological responses. *European Journal of Social Psychology*, *41*(6), 681-687.

Rosenberg, M. S. (2005). The file-drawer problem revisited: a general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution*, *59*(2), 464-468.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, *86*(3), 638.

Royal College of Nursing (2009) *The recognition and assessment of acute pain in children: update of full guideline*, Clinical Practice Guideline, London.

Rupinski, M. T., & Dunlap, W. P. (1996). Approximating Pearson product-moment correlations from Kendall's tau and Spearman's rho. *Educational and psychological measurement*, *56*(3), 419-429.

Schade, J. G., Joyce, B. A., Gerkensmeyer, J., & Keck, J. F. (1996). Comparison of three preverbal scales for postoperative pain assessment in a diverse pediatric sample. *Journal of pain and symptom management, 12*(6), 348-359.

Schultz, A. A., Murphy, E., Morton, J., Stempel, A., Messenger-Rioux, C., & Bennett, K. (1999). Preverbal, early verbal pediatric pain scale (PEPPS):

development and early psychometric testing. *Journal of pediatric nursing, 14*(1), 19-27.

Sikorova, L., & Hrazdilova, P. (2011). The effect of psychological intervention on perceived pain in children undergoing venipuncture. *Biomedical Papers of the Medical Faculty of Palacky University in Olomouc*, *155*(2).

Singer, A. J., Gulla, J., & Thode Jr, H. C. (2002). Parents and practitioners are poor judges of young children's pain severity. *Academic emergency medicine*, *9*(6), 609-612.

Soetenga, D., Frank, J., & Pellino, T. A. (1999). Assessment of the validity and reliability of the University of Wisconsin Children's Hospital Pain scale for Preverbal and Nonverbal Children. *Pediatric nursing, 25*(6), 670.

Solodiuk, J., & Curley, M. A. (2003). Pain assessment in nonverbal children with severe cognitive impairments: the Individualized Numeric Rating Scale (INRS). *Journal of Pediatric Nursing, 18*(4), 295-299.

Splinter, W. M., Semelhago, L. C., & Chou, S. (1994, February). The reliability and validity of a modified CHEOPS pain score. *Anesthesia and Analgesia.* . *78*(2), U220-U220.

Stallard, P., Williams, L., Velleman, R., Lenton, S., McGrath, P. J., & Taylor, G. (2002). The development and evaluation of the pain indicator for communicatively impaired children (PICIC). *Pain, 98*(1-2), 145-149.

Stewart, B., Lancaster, G., Lawson, J., Williams, K., & Daly, J. (2004). Validation of the Alder Hey triage pain score. *Archives of disease in childhood, 89*(7), 625-630.

Stinson, J., Yamada, J., Dickson, A., Lamba, J., & Stevens, B. (2008). Review of systematic reviews on acute procedural pain in children in the hospital setting. *Pain Research and Management*, *13*(1), 51-57.

St‐Laurent‐Gagnon, T., Bernard‐Bonnin, A. C., & Villeneuve, E. (1999). Pain evaluation in preschool children and by their parents. *Acta Paediatrica, 88*(4), 422-427.

Taddio, A., Nulman, I., Koren, B. S., Stevens, B., & Koren, G. (1995). A revised measure of acute pain in infants. *Journal of pain and symptom management, 10*(6), 456-463.

Tan, O. T., & Stafford, T. J. (1992). EMLA for laser treatment of portwine stains in children. *Lasers in surgery and medicine*, *12*(5), 543-548.

Tarbell, S. E., Cohen, I. T., & Marsh, J. L. (1992). The Toddler-Preschooler Postoperative Pain Scale: an observational scale measuring postoperative pain in children aged 1–5. Preliminary report*. Pain, 50*(3), 273-280.

Tyler, D. C., Tu, A., Douthit, J., & Chapman, C. R. (1993). Toward validation of pain measurement tools for children: a pilot study. *Pain*, *52*(3), 301-309.

Verghese, S. T., & Hannallah, R. S. (2010). Acute pain management in children. *Journal of pain research*, *3*, 105.

Vessey, J. A., Carlson, K. L., & McGill, J. (1994). Use of distraction with children during an acute pain experience. *Nursing research*.

Voepel-Lewis, T., Malviya, S., & Tait, A. R. (2005). Validity of parent ratings as proxy measures of pain in children with cognitive impairment. *Pain Management Nursing*, *6*(4), 168-174.

Voepel-Lewis, T., Shayevitz, J. R., & Malviya, S. (1997). for Scoring Postoperative Pain in Young Children. *Pediatric nursing*, *23*(3).

Von Baeyer, C. L. (2006). Children's self-reports of pain intensity: scale selection, limitations and interpretation. *Pain Research and Management, 11*(3), 157-162.

Von Baeyer, C. L., & Spagrud, L. J. (2007). Systematic review of observational (behavioral) measures of pain for children and adolescents aged 3 to 18 years. *Pain*, *127*(1-2), 140-150.

Williams, A. C. D. C., Davies, H. T. O., & Chadury, Y. (2000). Simple pain rating scales hide complex idiosyncratic meanings. *Pain*, *85*(3), 457-463.

Wolf, A. R., Hughes, D., Wade, A., Mather, S. J., & Prys-Roberts, C. (1990). Postoperative analgesia after paediatric orchidopexy: evaluation of a bupivacaine-morphine mixture. *British Journal of Anaesthesia, 64*(4), 430-435.

Wong, D. L., & Baker, C. M. (1988). Pain in children: comparison of assessment scales. *Pediatr Nurs*, *14*(1), 9-17.

Wright, G. Z., Weinberger, S. J., Marti, R. & Plotzke, O. (1991). The effectiveness of infiltration anesthesia in the mandibular primary molar region. *Ped Dent 13*, 278-282.

Yeh, C. H. (2005). Development and validation of the Asian version of the oucher: a pain intensity scale for children. *The Journal of Pain*, *6*(8), 526-534.

# CHAPTER TWO, EMPIRICAL PAPER:

# DEVELOPING A SCREENING TOOL TO DETECT GASTRIC PAIN IN CHILDREN WITH MINIMAL VERBAL COMMUNICATION

## 2.1. Abstract

**Background:** Gastro-oesophageal Reflux Disease (GORD) is a painful treatable health condition with an increased prevalence in people with Intellectual Disabilities (ID). GORD may be underdiagnosed in people with ID due to difficulties in self-reporting of pain, which is a primary symptom, and the invasive procedures required to confirm diagnosis. The Gastric Distress Questionnaire (GDQ) is a parent report questionnaire designed to screen for GORD in people who cannot self-report. The studies reported here offer an exploration of the features and validity of the GDQ and attempt to develop an accompanying brief observational tool.

**Method:** GDQ scores were compared to parent report of recent GORD for 599 children aged 1-18 with ID with and without a known underlying genetic syndrome and autism. Behavioural coding was conducted of footage of 49 children with ID. Observers coded behaviours from the GDQ which could be seen in brief observation periods.

**Results:** A five factor structure was established for the GDQ. Significant differences were found in four of the factor scores and the GDQ total between children with and without recent GORD. No significant relationship was found between behaviours recorded by naïve observers and GDQ scores provided by parents.

**Conclusion:** The GDQ may be a useful tool for detecting children who could benefit from medical investigation of GORD. Further study is required comparing GDQ scores to the outcomes of medical diagnostic procedures to establish construct validity. Furthermore, the findings reported here highlight the importance of parent report in recognising behavioural indicators of gastric pain in this population.

## 2.2. Introduction

In 2006 the United Nations published their convention on the rights of people with disabilities (CRPD), stating that all people should have *“the right to the enjoyment of the highest attainable standard of health without discrimination on the basis of disability”* (United Nations General Assembly, 2006). The UN CRPD was written not only to protect the rights of people with physical disabilities, but also those with intellectual disabilities (ID) of whom there are approximately 1.4 million living in the UK (Mencap, n.d.). The recent ‘Long Term Plan’ of the NHS outlines their commitment to tackle health inequalities, including providing the “right care” for children with ID (National Health Service, 2018). The inclusion of this pledge highlights the current limitations of UK health service provision for people with ID. These legislative and policy papers demonstrate the ongoing need to eliminate healthcare inequalities for people with ID.

Health inequalities exist in many forms and have a far-reaching impact on the lives of people with ID. People with ID often already face vulnerabilities, in some cases due to pathologies associated with the genetic syndromes which underlie some IDs. Emerson and Baines (2011) also highlight deficiencies in healthcare provision as playing a key role in the differences in physical health outcomes for people with ID when compared to the typically developing population. The life expectancy of a person with ID is 19.7 years lower than that of a typically developing individual (Glover, Williams, Heslop, Oyinlola & Grey, 2017). A relationship has been demonstrated between severity of ID and mortality rates, with people with profound or multiple ID having a life expectancy 20 years lower than that of someone with mild

ID (Heslop, Blair, Fleming, Hoghton, Marriott & Russ, 2013). Research has identified a range of specific health problems that occur more frequently in people with ID, including epilepsy, sensory impairments, digestive problems, reflux, respiratory disease, poor oral health, and periodontal disease (Anders & Davis, 2010; Emerson & Baines, 2011). Critically, 98% of people with ID who die prematurely have one or more known long term medical condition at the time of their death; 20% have seven or more known medical conditions (Heslop, Blair, Fleming, Hoghton, Marriott & Russ, 2013). Despite these and other data demonstrating the poor physical health outcomes associated with ID, people with ID often face barriers to recognition, diagnosis and treatment of health problems (Morin, Mérineau-Côté, Ouellette-Kuntz, Tassé & Kerr, 2012). Problems in communication are cited as one of the key barriers to treatment, with 70% of GPs reporting that they do not know how seriously to take health complaints made by people with ID (Lennox, Diggens & Ugoni, 1997). Other studies report a lack of training, appropriate assessment tools and clinician confidence as obstacles to improving outcomes (Malviya, Voepel-Lewis, Merkel & Tait, 2005; Lewis Gaffney & Wilson, 2017). As such, these studies demonstrate that a reduction in health inequality for people with ID is predicated on substantive changes to current practice, including better training and more specific tools to support clinicians working with this population.

Many of the health problems that are common in people with ID are known to cause physical pain. In typically developing populations, self-report of frequency, intensity, or location of pain is the gold-standard of measurement, even in children as young as three years old (Stinson, Kavanagh, Yamada, Gill & Stevens, 2006). However, many people with ID are unable to self-report. For example, 52% of people with ID who died prematurely had limited or no verbal communication (Heslop et al., 2013). Even in individuals with ID who are verbal, many struggle to effectively describe the nature and location of their pain (Findlay, Williams & Scior, 2014). Failing to detect pain in people with ID can have significant consequences in addition to failing to treat potentially treatable health conditions. Physical pain is strongly associated with behaviours that challenge and poor sleep (Carr & Owen-DeSchryver, 2007; Wiggs & Stores, 1996). The presence of behaviours that challenge increases the likelihood of reactive and restrictive behaviour management strategies, such as restraint or seclusion (Allen, Lowe, Brophy & Moore, 2009), and reduces quality of life (Beadle-Brown, Murphy & DiTerlizzi, 2009). The comorbidities associated with pain that have been demonstrated in people with ID provide further rationale for the investment in effective identification of pain and painful health conditions in this population.

There are tools available which can reliably identify pain in people who cannot self-report, the meta-analysed data in chapter one demonstrates that observational measures of pain correlate well to self-reported pain scores.  Tools such as the Face Legs Activity Crying Consobility scale (FLACC; Merkel, Voepel-Lewis, Shayevitz & Malviya, 1997) and Non-Communicating Child Pain Checklist (NCCPC; Breau et al, 2000) rely on observations of pain related behaviour from a care-giver or clinician and have shown good validity and reliability in measuring pain in people with ID (Crosta, Ward, Walkers & Peter, 2014). Although detection of pain in people with ID is an important step towards improving physical health outcomes, there is still a significant difference between recognising that someone is in pain and being able to accurately identify and treat the underlying causes. Measures that focus on identification of *specific* painful health conditions may offer clinical utility in aiding diagnosis and therefore increasing the likelihood of treatment. One health condition which may benefit from the development of such a measure is Gastro-oesophageal Reflux Disease (GORD). GORD occurs when stomach acid repeatedly returns to the oesophagus, resulting in pain to the throat and chest. GORD is easily treated in the majority of cases; however, left untreated it can result in permanent damage to the cells which line the oesophagus, a condition known as Barrett's oesophagus. Research suggests that people with Barrett's oesophagus may be up to ten times more likely to develop oesophageal cancer (Solaymani-Dodaran, Logan, West, Card & Coupland, 2004). People with ID are at disproportionately high risk of experiencing GORD, with some studies suggesting a prevalence in this population as high as 50% (Bohmer et al., 2000; Haveman, Heller, Lee, Maaskant, Shooshtari & Strydom, 2010). When GORD is present, people with ID have significantly higher rates of self-injurious behaviours and sleep problems (Luzzani, Macchini, Valade, Milani & Selicorni, 2003). However, GORD is also likely under-diagnosed in people with ID, because initial identification of the disease is typically based upon self-report of epigastric pain or heartburn (Hassal, 2001). Further challenges to successful treatment of GORD are conferred by the diagnostic assessment process; if GORD is suspected then painful and invasive procedures such as endoscopy are used to confirm the diagnosis (NICE, 2015) which clinicians may be hesitant to perform without significant justification of need. As such, NICE guidelines have cited the identification of behavioural markers of GORD as a current research priority (NICE, 2015). Thus, the development of a behavioural screening tool to aid identification of this common and under-recognised painful health condition in people with ID could have significant impact on unequal health outcomes.

In summary, health inequalities are well evidenced as affecting the duration and quality of life of people with ID. In particular, pain is common and can have a significant impact of behaviour and quality of life, but is hard to recognise due to the communication impairments that are prevalent in this population. There are particular health problems which affect people with ID disproportionately in comparison to the typically developing population, and improving recognition and diagnosis of such health problems may be one step towards tackling the broader issue of health inequality. GORD is a health problem which is more common is people with ID, and is of particular interest because it is painful and can have long term consequences if left untreated, but is frequently easily treatable if diagnosed.

Therefore, the current study investigates a tool which shows promise in the identification of gastric pain and GORD symptoms in people who cannot verbally report their internal experiences. The Gastric Distress Questionnaire (GDQ) is a caregiver report questionnaire which asks about the frequency of observable behaviours related to GORD (Oliver & Wilkie, 2005). There has been limited exploration of the psychometric properties of the GDQ. Clinical utility of the GDQ would be improved if the factors underlying the measure were understood, and if a cut-off was established that could identify children who may benefit from further investigation for GORD symptoms. The most robust method to study the validity of a measure of GORD would be to compare the measure against gold-standard clinical diagnostic procedures. However, medical diagnosis typically involves painful invasive procedures such as endoscopy, which would pose significant ethical considerations. Most children with ID have continuous support from either a parent or other caregiver who would be well placed to reliably report on both recent behaviours and any current or previous diagnosis of GORD. Although the GDQ is not an age specific measure, focusing an initial investigation on children provides an opportunity to explore the relationship between scores on the GDQ and diagnosis of GORD, using the knowledge of caregivers, avoiding the need for invasive medical procedures. If a relationship is found between GDQ scores and parent/caregiver report of GORD then this would offer support for the clinical utility of the measure and provide justification to conduct further investigation utilising medical confirmation of diagnosis.

A second opportunity is to examine the utility of the behavioural indicators in the GDQ as a brief observational screening tool for naïve observers or clinicians. The GDQ is reliant on caregiver report; however, not everyone with ID has access to someone who could reliably report on their behaviours. Additionally, studies have indicated that clinicians often express uncertainty regarding the accuracy of caregiver reports in medical settings (Lewis, Gaffney & Wilson, 2017). Given that many clinicians feel ill equipped to assess pain and physical health people with ID themselves, development of an appropriate tool to support clinical judgement may help address this barrier. One investigation suggests that as many as one in five problems in diagnosis were directly related to issues in accessing specialist care, including referrals not being made (Heslop, Blair, Fleming, Hoghton, Marriott & Russ, 2013). Providing a screening tool for primary care settings which could help to identify those people that would benefit from referral to specialist services, would help to ensure that more people with ID could receive appropriate diagnosis and treatment.

In conclusion, the studies reported here address two broad aims:

1. To provide a preliminary investigation of the structure, sensitivity, and specificity of the GDQ
2. To explore the potential feasibility of converting the GDQ into a brief observational screening tool for use by naïve observers in primary care settings.

## 2.3. Study one

Study one aims to provide a preliminary investigation of the psychometric properties of the GDQ by addressing the following aims:

1. To explore the underlying factor structure of the GDQ
2. To explore to what extent GDQ scores distinguish children with GORD from those without GORD, according to parent/caregiver reports of GORD diagnosis
3. To identify a clinical cut-off score for the GDQ which would provide adequate sensitivity and specificity for use as a screening tool in primary care settings
4. To explore whether GDQ scores relate specifically to symptoms of GORD rather than pain caused by other underlying health conditions.

### 2.3.1 Method

This study utilised data collected from previous studies carried out by the Cerebra Centre for Neurodevelopmental Disorders. In order to collect a heterogenous and representative sample of children with ID, data were taken from studies investigating multiple different genetic syndromes associated with ID as well as children with ID without a known genetic syndrome, and children with autism. In all of the included studies the GDQ, the Wessex, and a background health questionnaire were included in the protocol and completed at the same time by parents/caregivers of children with ID. Recruitment was conducted through syndrome specific charities, parent support groups, and schools, residential, and day services. More specific information regarding recruitment is available from Richards, Oliver, Nelson and Moss (2012), Arron, Oliver, Moss, Berg and Burbidge (2011), and Oliver, Berg, Moss, Arron and Burbidge (2011).

#### 2.3.1.1. Measures.

*Gastric Distress Questionnaire* (GDQ; Oliver & Wilkie, 2005) The GDQ is an informant report measure comprising 17 items reporting on behaviours observed in the previous two weeks. The majority of the behaviours are scored on a five point likert scale relating to the frequency with which the behaviour has been observed, from “not occurred” to “occurs more than once an hour”.

*Health Questionnaire:* (Hall, Arron, Sloneem & Oliver, 2008) The health questionnaire is a background questionnaire used to collect information regarding health problems that the child has been diagnosed with and/or treated for both historically and recently. For the purposes of the current study, data regarding the severity of health problems reported in the last month were utilised.

*The Wessex*: (Kushlick, Blunden & Cox, 1973) The Wessex assesses physical and social abilities via caregiver report. It scores across five domains; continence, mobility, self-help skills, speech, and literacy. Kushlick, Blunden & Cox (1973) report that the scale has good inter-rater reliability for both children and adults with ID.

#### 2.3.1.2. Sample.

The databases of several studies which included the GDQ and Health Questionnaire were included. To be included in the analysis participants were required to meet the following inclusion criteria:

1. Older than one year of age
2. Not older than 18 years of age
3. Complete data for all items on the GDQ
4. Valid response to the question "Has your child experienced gastric reflux in the last month?" from the Health Questionnaire

Due to the inclusion of multiple studies, nine cases of duplicate data were identified. In these cases, only the earliest valid data set collected for each individual was included in the analysis. Before any further analysis was conducted, the GDQ total was checked for outlying scores using visual inspection of a basic box and whisker plot. Two participants with outlying scores were removed.

Table 17 displays the details of the final sample. After removal of participants who did not meet inclusion criteria, duplicates, and outliers, the final sample was *N* = 599 with a mean age of 9.37 (range 1 - 18) years. A total of 13 known genetic syndromes were included in the study, as well as children with a primary diagnosis of autism and children with a primary diagnosis of ID with no known cause. For the majority of the sample (*n* = 560) there were also data available from the Wessex (Kushlick, Blunden & Cox, 1973) relating to functional behaviours and sensory impairments which is outlined in Table 17.

#### 2.3.1.3. Analysis.

The primary outcome measure was GDQ total scored as a sum of the likert values of all 17 items. This was assessed against the parental report of gastric reflux in the previous month. Data on reflux from the Health Questionnaire were collected according to severity – none, mild, moderate or severe, however the numbers for each severity group were small, so to maximize sample size and power, the groups were condensed into a binary outcome (no reported reflux/ reported reflux), in which 'reported reflux' included all levels of severity.  A Shapiro-Wilk analysis on the GDQ total was significant, indicating skewness within the sample, therefore non-parametric analyses were used throughout.

The first aim of the present study was to establish the underlying structure of the GDQ and examine the existence of any factors. Principal components analysis (PCA) with varimax rotation was utilised, in accordance with other studies reporting exploratory factor analysis of pain measures (Hermann, Zohsel, Hohmeister & Flor, 2008; von Baeyer, Chambers & Eakins, 2011). Consistent with previous studies, factors with eigenvalues > 1 were retained, and item loading was determined based on the selection of items with loading values > .4 in the varimax rotation (Field, 2005).

*Table 17:* Age, sex, diagnosis, functional behaviour, and sensory impairments relating to the sample of study one (*N* = 599).

| | ***n*** | **% of sample** |
|---|---|---|
| Age group | | |
| 1-3 years | 52 | 8.68 |
| 4-7 years | 169 | 28.21 |
| 8-11 years | 179 | 29.88 |
| 12-15 years | 137 | 22.87 |
| 16-18 years | 62 | 10.35 |
| Sex | | |
| Male | 424 | 70.80 |
| Female | 175 | 29.20 |
| Diagnosis | | |
| Angleman Syndrome | 13 | 2.17 |
| Cri du Chat Syndrome | 24 | 4.01 |
| Cornelia de Lange Syndrome | 32 | 5.34 |
| Fragile X Syndrome | 65 | 10.85 |
| Prader Willi Syndrome | 60 | 10.02 |
| Lowe Syndrome | 23 | 3.84 |
| Smith Magenis Syndrome | 8 | 1.34 |
| Soto Syndrome | 22 | 3.67 |
| Tuberous Sclerosis Complex | 45 | 7.51 |
| Phelan McDermid Syndrome | 29 | 4.84 |
| 1p36 | 18 | 3.01 |
| 9q34 | 11 | 1.84 |
| 8p23 | 15 | 2.50 |
| Autism | 222 | 37.06 |
| ID no known cause | 12 | 2.00 |
| Sensory impairments | | |
| Poor hearing/deaf (n=558) | 39 | 6.50 |
| Poor vision/blind (n=556) | 90 | 15.00 |
| Non-verbal (n=559) | 86 | 14.40 |
| Self-Help (n=560) | | |
| Not able | 151 | 25.20 |
| Partly able | 224 | 37.40 |
| Able | 185 | 30.90 |
| Mobility (n=560) | | |
| Non-ambulant | 58 | 9.70 |
| Partly-ambulant | 64 | 10.60 |
| Ambulant | 438 | 72.80 |

Comparisons of the GDQ score between children with and without GORD were carried out using a Mann-Whitney U analysis. The relationship between GORD and the identified factors was explored using a t-test applied to regression factor scores. To explore sensitivity and specificity and inform the choice of a clinical cut-off, a Receiver Operating Charactistics (ROC) analysis was conducted. Youden's Index (YI) was calculated for each of the potential cut-off values. YI gives a metric between zero and one indicating the balance between sensitivity and specificity of a measure (Fluss, Faraggi & Reiser, 2005; Liu 2012). A YI of  one would indicate a measure which successfully identified every person with GORD without wrongly mislabelling any individual without GORD as having it.

Finally, a series of Chi$^2$ analyses were conducted to explore associations between the proposed clinical cut-off and the presence of painful health problems, including GORD. To account for multiple analyses the alpha value was adjusted to $\alpha = .001$. Odds Ratios (OR) were calculated for any health problems that returned a significant results in the Chi$^2$ analysis.

## 2.3.2. Results

### 2.3.2.1. Analysis of GDQ factor structure.

To achieve the first aim of exploring the structure of the GDQ, analyses were carried out on the individual GDQ items. Bartlett's test of sphericity was found to be significant ($X^2$ = 2007.48, $p < .001$) supporting the hypothesis that the items fit an identity matrix making the GDQ amenable to factor analysis. The high value found in the Kaiser-Meyer-Olkin test confirm the sample as being sufficient for factor analysis (KMO = .84). The PCA identified a five-factor solution, seen in Table 18, which accounted for 54.36% of the variance[5]. All 17 of the GDQ items loaded on to at least one factor. One GDQ item, item 11 which asks how often the child cries, moans or otherwise appears to be in pain, loaded across multiple factors, contributing to both factors one and three.

[5] Although 60% is often cited as the minimum level of variance to consider factor analysis acceptable, it is not uncommon to consider models which explain 50-60% in the social sciences (Hair, 2014).

*Table 18*: Five factor solution produced by PCA with varimax rotation, including contribution of each factor to the 54.36% of variance explained by the model.

| | **Factor Loadings** | | | | |
|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** |
| Item 5. Place their hands or fingers in back of their mouth | .77 | | | | |
| Item 6. Chew on his/her clothes, fingers, hands or other parts of the body, objects or material | .66 | | | | |
| Item 3. Salivate excessively | .63 | | | | |
| Item 7. Grind their teeth | .54 | | | | |
| Item 8. Scratch/hit/press/rub around the upper chest/throat | .52 | | | | |
| Item 10. Cough, gag or regurgitate | .51 | | | | |
| Item 2. Lie over an object on his/her stomach | | .82 | | | |
| Item 1. Arch his/her back, lie over arms of chairs or people | | .75 | | | |
| Item 4. Fidget, wriggle or move their body a great deal | | .58 | | | |
| Item 13. Appear indecisive about food | | | .80 | | |
| Item 12. Refuse food even though they are probably hungry | | | .79 | | |
| Item 11. Appear in pain or discomfort (cry, groan or moan) | .41 | | .45 | | |
| Item 9. Drink/request/seek out an excessive amount of fluids | | | | .73 | |
| Item 16. Bad breath | | | | .64 | |
| Item 14. Wake during the night | | | | .41 | |
| Item 17. Experience frequent respiratory infections | | | | | .77 |
| Item 15. Sleep sitting or propped up | | | | | .57 |
| Total contribution of component to model (%) | 15.89 | 11.46 | 10.55 | 8.83 | 7.64 |

### 2.3.2.2. Analysis of validity.

To explore the second aim to explore the GDQ's ability to identify children who may be experiencing GORD, the total score and factor level scores on the GDQ were compared to parental rating of gastric reflux in the last month from the Health Questionnaire. The data presented in Table 19 demonstrate that children with caregiver reported GORD had significantly higher total scores on the GDQ and significantly higher scores on Factors one, three, four, and five of the GDQ. The comparison on Factor two GDQ scores for those with and without GORD was non-significant.

*Table 19:* Median (IQR) of GDQ total and regression factor scores for those with and without recent GORD according to caregiver report, with associated Mann-Whitney U / t-test statistics and *p* values for between group comparisons.

| Score | Median scores (IQ range)<br>With GORD | Without GORD | Mann-Whitney U / *t* value | *p* |
|---|---|---|---|---|
| GDQ total | 19.50 (13.00) | 11.00 (14.00) | 15541.00 | < .001 |
| Factor 1 | 0.36 (1.87) | -0.36 (0.95) | 5.25 | <. 001 |
| Factor 2 | -0.23 (1.56) | -0.19 (1.50) | -.21 | .835 |
| Factor 3 | 0.07 (1.80) | -0.37 (0.68) | 4.66 | < .001 |
| Factor 4 | 0.15 (1.53) | -0.30 (1.38) | 3.10 | .002 |
| Factor 5 | -0.06 (1.94) | -0.34 (0.79) | 3.26 | .001 |

#### 2.3.2.3. Sensitivity and specificity of GDQ cut-off values.

In order to answer the third research question and establish a suitable clinical cut-off for the GDQ, a ROC analysis was conducted. Given the results of the PCA, the ROC analysis in Figure 8 included both the full total of the GDQ as a sum of all item vales, as well as an adjusted total which omitted the items contributing to factor two which was demonstrated to be non-significant (see Table 18, above). Area under the curve was 0.73 (*p* < .001) for the GDQ total score and 0.75 (*p* < .001) for the adjusted score. Results from the Youden's Index analysis are presented in Table 20. The best YI achieved was YI = .39, using the adjusted GDQ total with cut off of > 8 where a score of nine or more is seen as indicative of potential GORD. This equates to a sensitivity of .83, and specificity of .56. In all the following analyses, the cut off of > 8 is referred to as the "clinical cut off".

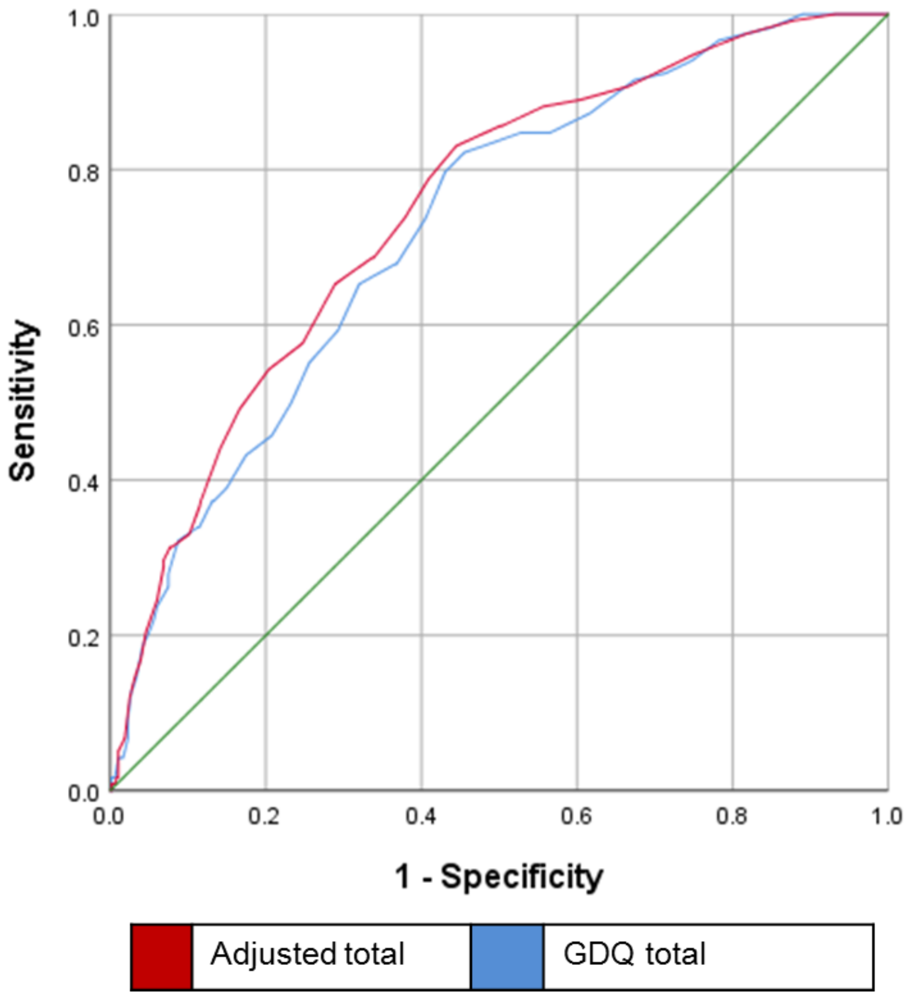


*Figure 8*: ROC curve for GDQ total and adjusted total calculated by omitting items one, two and four from the total.

*Table 20:* Sensitivity, specificity, and Youden's Index scores for potential cut off values on the GDQ with all items summed, and the adjusted sum with items one, two, and four omitted

| GDQ total | | | | GDQ adjusted total | | | |
|---|---|---|---|---|---|---|---|
| Cut off value | Sensitivity | Specificity | Youden's Index | Cut off value | Sensitivity | Specificity | Youden's Index |
| > 10 | .85 | .47 | .32 | > 7 | .86 | .50 | .36 |
| > 11 | .83 | .52 | .35 | > 8 | .83 | .56 | .39 |
| > 12 | .82 | .55 | .37 | > 9 | .79 | .59 | .38 |
| > 13 | .80 | .57 | .37 | > 10 | .74 | .62 | .36 |
| > 14 | .74 | .60 | .33 | > 11 | .69 | .66 | .35 |

#### 2.3.2.4. GDQ associations with other painful health problems.

The final research question was addressed with a series of Chi$^2$ and odds ratio (OR) analyses to examine whether the GDQ score associated with GORD specifically or whether it may also be associated with other health problems experienced by the child in the previous month. Using the clinical cut-off of > 8 recommended by the YI (see Table 20, above), participants were grouped into those who scored above and those who scored below clinical cut off. The presence of physical health problems in these two groups was then compared.

The results in Table 21 show that those with parent reported GORD, bowel problems, epilepsy, respiratory problems, and skin problems were significantly more likely to score above the cut-off on the GDQ. No significant differences were found between children above and below GDQ cut-off in the rates of diabetes, liver or kidney problems, ear problems, or heart problems. The OR calculations show that the odds of a child having parent reported GORD if they score above cut off on the GDQ are six times the odds of a child having parent reported GORD if they scored below cut off. The OR was slightly lower for bowel problems and respiratory problems (OR = 5.05 and 4.84 respectively), and considerably lower for epilepsy (OR = 2.19), and skin problems (OR = 1.99).

*Table 21*: $X^2$ and Odds Ratio analysis of health problems in the last month, according to parental report, in children scoring above and below clinical cut off on the GDQ

| **Parent Reported Health problem** | ***n*** | **Above GDQ cut off** | | **Below GDQ cut off** | | ***X²*** | ***p*** | **OR [95% CI]** |
|---|---|---|---|---|---|---|---|---|
| | | **With problem** | **Without problem** | **With problem** | **Without problem** | | | |
| GORD | 599 | 98 | 214 | 20 | 267 | 56.45 | < .001 | 6.11 [3.66, 10.22] |
| Bowel problems | 597 | 102 | 210 | 25 | 260 | 50.89 | < .001 | 5.05 [3.15, 8.11] |
| Respiratory problems[1] | 599 | 50 | 262 | 10 | 277 | 26.09 | < .001 | 4.84 [2.51, 9.31] |
| Ear problems | 597 | 47 | 263 | 21 | 266 | 9.09 | < .01 | 2.26 [1.32, 3.89] |
| Epilepsy | 598 | 60 | 252 | 28 | 258 | 10.60 | .001 | 2.19 [1.36, 3.55] |
| Skin problems[2] | 595 | 130 | 178 | 76 | 211 | 16.23 | < .001 | 1.99 [1.41, 2.80] |
| Heart problems | 597 | 14 | 297 | 7 | 279 | 1.85 | .17 | - |
| Diabetes | 595 | 8 | 301 | 2 | 284 | 3.21 | .07 | - |
| Liver/Kidney problems | 599 | 20 | 292 | 9 | 278 | 3.48 | .06 | - |

[1] GDQ question 17 removed due to potential confound

[2] GDQ question 8 removed due to potential confound

## 2.4. Study Two

Study Two aimed to explore the feasibility of a brief observational screening tool for use in clinical settings by addressing the following aims:

1. To explore whether behaviours reported by parents/caregivers on the GDQ can be detected in brief observation by a naïve observer
2. To evaluate the degree of association between GDQ scores obtained from parent / caregiver report and observation by a naïve observer
3. To explore whether scores obtained from brief observation by a naïve observer can differentiate children whose GDQ scores fall above and below the clinical cut-off derived in Study One.

### 2.4.1. Method

Participants from previous Cerebra Centre studies for whom both video footage and GDQ data were collected were identified for potential inclusion in the study. The footage reviewed had never previously been investigated for behaviours indicative of GORD. The observers responsible for rating the behaviours in the current study had no involvement in the original studies and had no interaction with the children or families and therefore were naïve observers. The observers were blinded to the GDQ scores until all videos had been coded.

For each child, ten minutes of video footage was selected for coding. Some of the studies had recorded several different naturalistic and experimental paradigms, therefore a selection hierarchy was developed to guide footage selection. The hierarchy aimed to capture naturalistic behaviour and, where possible, avoid capturing the effects of experimental manipulation on behaviours. The hierarchy also sought to avoid inclusion of distress that may be directly induced by task demands or biased selection of footage which showed disproportionate levels of distress. As such, the obtained video clips were considered to be representative of children's typical behaviour.

The different experimental paradigms included in the original footage are briefly described below. The experimental paradigms and recruitment strategies are reported in greater detail in the original studies (Arron et al., 2011; Oliver et al., 2011; Richards et al., 2012).

*Autism Diagnostic Observation Schedule*: (ADOS; Gotham, Risi, Pickles & Lord, 2006) The ADOS is a behavioural assessment used to elicit typical social behaviours in order to investigate potential autism related behaviours. The ADOS includes a minimum of two minutes of free-play, and the prescribed sections of the assessment are play activities for the children. The free-play was included in the selected footage for all participants for whom ADOS footage was used. ADOS footage was used for 18% of participants (*n* = 9).

*Social presses*: The social presses is a series of play sessions with some scripted tasks such as tower building and ball throwing. Included in the current study were the 'Responsive Engagement' condition, in which the adult in the play session responds appropriately and naturalistically to any interaction initiated by the child, and the 'Active Engagement' condition, in which the adult actively engages the child in play. Conditions in which the adult ignores interactions from the child, or places explicit task demands on the child, were excluded due to the potential for distress. Social presses footage was used for 53% of participants (*n* = 26).

*Experimental functional analysis analogues*: The EFA analogues are a series of social interactions, designed to examine a child's response to social engagement, task demand, and being ignored. There are no toys provided or tasks presented to the child. Only the 'high attention' control condition was used for the current study as this was thought to be the condition least likely to elicit distress for the majority of children. During the 'high attention' condition, the adult actively engages with the child without placing demands on them, and responds naturalistically to any child initiated interactions. Footage taken from EFA analogues was used for 20% of participants (*n* = 10).

*Naturalistic observation*: Naturalistic observation footage describes footage taken of the children without any experimental manipulation, e.g. footage of the child playing at home, or classroom footage of a typical lesson. Naturalistic footage was the preferred footage for inclusion in the current study, but because it could only be used if the original study had included the routine collection of naturalistic footage in its protocol, it was the least frequently available footage. Footage of natural observations was used for 8% of participants (*n* = 4).

#### 2.4.1.1. Sample.

After the coding scheme was established, the following inclusion criteria were applied to the available videos:

1. Child aged between one and 18 years

2. A minimum of ten minutes of video footage available
3. Child must be visible on screen for at least 90% of the duration of the footage
4. A GDQ must have been completed by the parent/carer within one month of the video footage capture

A total of 31 videos were excluded, the majority because the date of completion on the GDQ was more than a month from the date of the video recording. The final sample therefore included footage of *N* = 49 children whose demographic details are displayed in Table 22.

*Table 22:* Age, sex, and known diagnoses for the *N* = 49 children included in study Two.

| | ***n*** | **%** |
|---|---|---|
| *Age:* Mean in years<br>(range) | 9.94<br>(2.00 - 18.00) | - |
| Gender | | |
| Male | 25 | 51.00 |
| Female | 24 | 49.00 |
| Diagnosis | | |
| Angelman Syndrome | 14 | 28.60 |
| Cri du Chat | 14 | 28.60 |
| Cornelia de Lange Syndrome | 9 | 18.40 |
| Autism | 10 | 20.40 |
| Smith-Magenis Syndrome | 1 | 2.00 |
| Cerebral palsy | 1 | 2.00 |

#### 2.4.1.2. Coding.

An observational behavioural coding scheme was developed from the GDQ using an iterative process to identify a set of behavioural definitions relating to gastric distress that could reliably be coded by researchers. Observable behaviours reported in the GDQ were identified and operationalised. A sub-set of 25% of the 80 videos available before exclusions were selected and coded independently by the two naïve observers and inter-observer reliability was calculated. Items with low reliability were reviewed and the definitions were revised to better reflect the behaviours observed in the videos. Some behaviours were combined into one code due to the difficulty of accurately and reliably differentiating between similar behaviours in video footage, for example, putting fingers in back of mouth, and chewing on fingers/hands. The videos were then re-coded by both researchers used the final coding scheme, which produced a good level of reliability (mean kappa = .78, range .61 - 1.00). The final coding scheme is presented in Table 23.

*Table 23:* Final coding scheme with levels of agreement reached. Where kappa is not reported behaviour was not observed in any of the videos selected for analysis of inter-rater reliability.

| **Behaviour** | **GDQ item(s)** | **Frequency/ Duration** | **Operationalised description** | **Kappa** |
|---|---|---|---|---|
| Back arching | 1 | Duration | A significant movement bending backwards or pushing the chest forwards in a way that creates an arch in the back | .97 |
| Lying on object or person | 1 | Duration | Lying down with back bent over/on top of an object, item of furniture or person in such a way as to create an arch in the spine | 1.00 |
| Lying prone on object or person | 2 | Duration | Lying down with stomach bent over/on top of an object, item of furniture, or person in such a way as to create a bend in the spine | .91 |
| Salivation | 3 | Duration | Visible saliva e.g. saliva on face or spitting, duration code ends (i) when saliva on face is no longer visible for any reason (e.g. head turns away, saliva dried or wiped), (ii) saliva is still visible but not on face e.g. spitting saliva on to surface, code ends when spitting stops | .70 |
| Swallowing | 3 | Frequency | Visible swallow or audible gulping noise in the absence of drink/food | - |
| Repetitive hand movements / Hand fidgeting | 4 | Duration | Repetitive movements of the hands, arms, and/or fingers without an apparent function. Do not code where the movement appears to be an exaggeration or repetition of a communicative or functional action e.g. repeating a sign or clapping | .68 |
| Repetitive body movements / Body fidgeting | 4 | Duration | Repeated movements of the body (excluding hands and fingers, see hand fidgeting) which appear to lack a clear function. Do not code where the movement appears to be an exaggeration or repetition of a communicative or functional action, or where the action is a direct response to an external stimulus (e.g. jiggling about in response to being tickled) | .71 |
| Hands in mouth | 5 & 6 | Duration | Placing fingers or hands in the mouth, either with or without visible chewing occurring | .76 |
| Chewing on clothes or object | 6 | Duration | Placing any non-food item in the mouth e.g. clothes, toys | .94 |

| **Behaviour** | **GDQ item(s)** | **Frequency/ Duration** | **Operationalised description** | **Kappa** |
|---|---|---|---|---|
| Teeth grinding | 7 | Duration | Either one or both of: Visible grinding of teeth - jaws clenched with movement in lower jaw, audible grinding of teeth - audible scraping noise in the absence of any other obvious source | .87 |
| Chest contact | 8 | Duration | Intentional direct contact (including scratching, hitting, pressing, rubbing) using a body part or object to any area below the chin and above the ribs. Exclude where contact is an action of wider communication e.g. Makaton, BSL. | .78 |
| Drink | 9 | Frequency | Seeking out a drink or visibly drinking | - |
| Coughing | 10 | Frequency | Visible or audible cough | .75 |
| Gagging/Regurgitating | 10 | Frequency | choking or retching noise or visible difficulties in swallowing accompanied by a forward motion in the shoulders | - |
| Crying | 11 | Duration | Sobbing or crying with or without visible tears, with a facial expression indicating distress e.g. two lines on the forehead, a furrowed brow | .89 |
| Groaning/moaning | 11 | Duration | A sustained low pitch noise accompanied by a distressed facial expression e.g. grimace or furrowed brow | .61 |
| Direct communication of pain | 11 | Frequency | Saying pain associated word such as "ouch" or using Makaton sign or picture symbol to communicate pain to others. Only code Makaton sign where there is clear evidence that this is the intention of the movement. e.g. use of other signs, interpretation by others | .62 |

Behaviours were coded live using Obswin behavioural coding software (Martin, Oliver & Hall, 1998). Twelve of the 17 behaviours were coded as duration variables, meaning that the length of time that the child spent engaged in the behaviour was recorded, e.g. length of time spent lying prone on an object. Five behaviours were coded as frequency variables, meaning that only a frequency count for the behaviour was kept, e.g. recording the behaviour of coughing as the number of individual coughs, as opposed to length of time spent coughing.

#### 2.4.1.3. Analysis strategy.

Coding of the full sample of videos was conducted using OBSWIN software (Martin, Oliver & Hall, 1998). Ten minutes of footage was coded for each child. Initial exploratory analysis of the frequency of behaviours within the sample was conducted.

Further analysis was conducted using the total number of behaviours shown by each child, the total frequency of behaviours across all behaviour types, the number of seconds spent engaged in any behaviour, and the percentage of time spent engaged in each behaviour. The different observational scores were compared to parent/caregiver GDQ scores using Spearman Rho correlations and to groups above and below the GDQ clinical cut-off, calculated in study one, using Mann-Whitney U analyses.

### 2.4.2. Results

#### 2.4.2.1. Exploratory analysis of types and frequency of observed behaviours.

To explore whether the behaviours reported on in the GDQ can be detected by a naïve observer in a brief period of observation an initial exploratory analysis of the frequency of observed behaviours was conducted, displayed in Figure 9. The majority of children (*n* = 28) engaged in between two to four different target GDQ behaviours in the course of a ten minute observation period. Table 24 displays the rates at which each of the behaviours were observed in the sample. The most commonly observed behaviour was hand fidgeting which was displayed by more than half of the sample (*n* = 29). Swallowing, gagging and drinking behaviours were not displayed by any of the children observed.

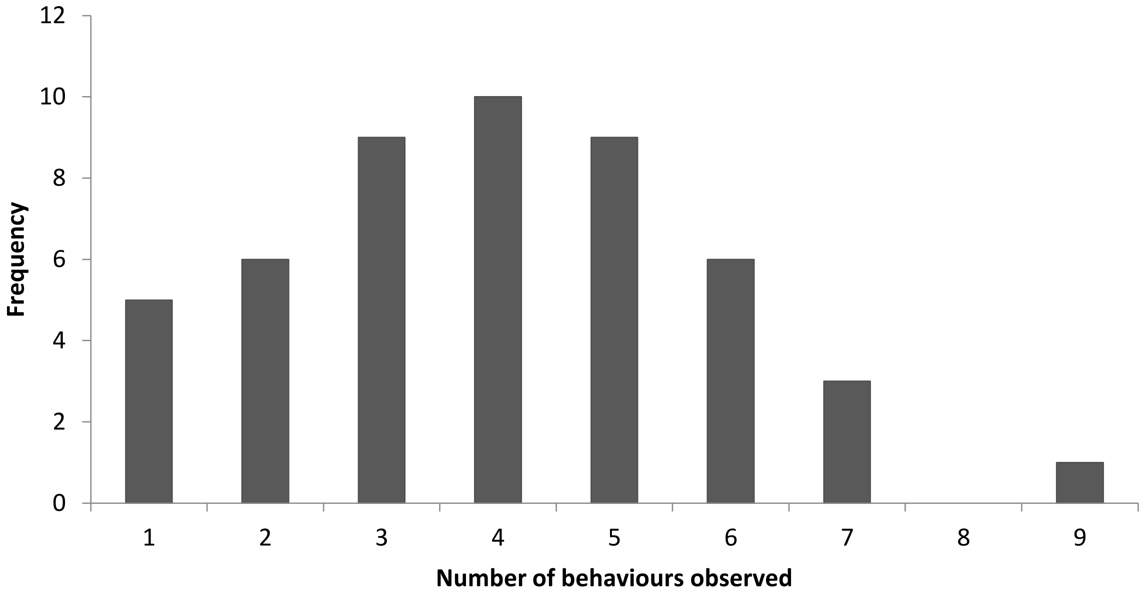


*Figure 9*: Histogram displaying the number of different target behaviours shown by each child

*Table 24*: Summary statistics associated with each observed behaviour; number of children observed engaging in behaviour; mean frequency of behaviour in 10 minute observation; total duration engaged in behaviour (s); duration engaged in behaviour as % of 10 minute observation period

| **Behaviour** | ***n*** | **Frequency of behaviour (Mean; range)** | **Total duration of behaviour (sec) (Mean; range)** | **Percentage duration of behaviour (%) (Mean; range)** |
|---|---|---|---|---|
| Repetitive hand movements / Hand fidgeting | 29 | 7.31<br>(1-41) | 23.37<br>(2-93) | 3.90<br>(0.33-15.50) |
| Repetitive body movements / Body fidgeting | 25 | 4.88<br>(1-13) | 23.80<br>(1-111) | 3.97<br>(0.17-18.47) |
| Chewing on clothes or object | 24 | 5.92<br>(1-20) | 88.69<br>(3-586) | 14.77<br>(0.50-97.50) |
| Hands in mouth | 18 | 5.44<br>(1-25) | 28.89<br>(3-154) | 4.19<br>(0.50-25.70) |
| Salivation | 14 | 3.79<br>(1-12) | 20.50<br>(2-102) | 3.41<br>(0.33-16.97) |
| Chest contact | 12 | 5.75<br>(1-26) | 16.92<br>(1-90) | 2.83<br>(0.17-14.98) |
| Coughing | 6 | 1.33<br>(1-3) | 1.50<br>(1-3) | .25<br>(0.17-0.50) |
| Crying | 6 | 5.67<br>(1-16) | 18.67<br>(2-49) | 2.95<br>(0.33-8.17) |
| Back arching | 4 | 4.75<br>(1-11) | 41<br>(2-133) | 6.83<br>(0.33-22.17) |
| Lying prone on object or person | 2 | 3.50<br>(1-6) | 6.00<br>(6-6) | 1<br>(1.00-1.00) |
| Groaning/moaning | 2 | 1.50<br>(1-2) | 3.50<br>(3-4) | 0.59<br>(0.50-0.67) |
| Direct communication of pain | 2 | 3.00<br>(2-4) | 3.00<br>(2-4) | 0.50<br>(0.33-0.67) |
| Lying on object or person | 1 | 1.00 | 13.00 | 2.17 |
| Teeth grinding | 1 | 1.00 | 3.00 | 0.50 |
| Swallowing | 0 | - | - | - |
| Drink | 0 | - | - | - |
| Gagging/Regurgitating | 0 | - | - | - |

#### 2.3.2.2. Comparison of observations against GDQ scores.

To address the second aim, the degree of association between GDQ scores obtained from parent / caregiver report and behaviours coded by a naïve observer was evaluated. The results of the Spearman's Rho correlations shown in Table 25 indicate that there were no significant associations between any of the coded behaviours and the adjusted GDQ total scores.

*Table 25*: Spearmans Rho correlations of adjusted GDQ score (items  one, two, and four ommitted) compared to total number of different types of behaviour recorded in ten minute observation; combined frequency of behaviours across all behaviour types; total duration engaged in any behaviour in seconds; total duration engaged in any behaviour as % of ten minute observation period

| | **Number of behaviours recorded** | **Combined frequency all behaviours** | **Total period engaged in any behaviour (sec)** | **Percentage period engaged in any behaviour (%)** |
|---|---|---|---|---|
| *r* | .10 | .04 | .08 | .08 |
| *p* | .48 | .79 | .58 | .58 |

Fidget behaviours were observed at very high frequency in behavioural observations, yet analysis in Study One (see Section 2.2.2) suggested that fidgeting behaviours were not associated with GORD. Therefore, to ensure that putative associations between observed behaviours and parent/caregiver GDQ were not obscured by the high frequency fidgeting behaviours, all correlations were replicated without hand and body fidgeting behaviours.  The results in Table 26 demonstrate that no associations were identified between the coded behaviours and the adjusted GDQ scores.

*Table 26*: Spearmans Rho correlations of adjusted GDQ score (items  one, two, and four ommitted) compared to total number of different types of behaviours, excluding hand and body fidgeting behaviours, recorded in ten minute observation; combined frequency of behaviours across all behaviour types; total duration engaged in any behaviour in seconds; total duration engaged in any behaviour as % of ten minute observation period

| | **Number of behaviours recorded** | **Combined frequency all behaviours** | **Total period engaged in any behaviour (sec)** | **Percentage period engaged in any behaviour (%)** |
|---|---|---|---|---|
| *r* | .11 | .05 | .05 | .05 |
| *p* | .48 | .75 | .71 | .76 |

### 2.3.2.3. Observed behaviours in children above and below GDQ cut-off.

Finally, to explore the potential association between coded behaviours and the GDQ cut-off a series of Mann Whitney U analyses were comparing coded behaviours between those children scoring above and below cut off on the GDQ. The results in Table 27 reveal that there were no significant associations between coded behaviours and GDQ clinical cut-off.

*Table 27*: Median, IQR, and Mann-Whitney U analysis grouped by children above and below clinical cut off on GDQ. Observation scores calculated as total number of different types of behaviour recorded in 10 minute observation; combined frequency of behaviours across all behaviour types; total duration engaged in any behaviour in seconds; total duration engaged in any behaviour as % of 10 minute observation period

| | **Median (IQR)** | | | |
|---|---|---|---|---|
| | **Below GDQ cut-off** | **Above GDQ cut-off** | **Mann-Whitney U** | ***p*** |
| Number of behaviours recorded | 2.00 (2.00) | 3.00 (2.25) | 133.50 | .07 |
| Combined frequency across all behaviours | 8.00 (29.00) | 13.00 (20.75) | 157.00 | .21 |
| Total period engaged in any behaviour (sec) | 18.00 (149.00) | 52.50 (156.75) | 163.00 | .27 |
| Percentage period engaged in any behaviour (%) | 3.00 (24.88) | 8.76 (26.13) | 162.50 | .27 |

## 2.5. Discussion

The studies reported here set out to explore and the utility of the GDQ as a tool to screen for GORD symptoms in children with ID. Study One aimed to explore the underlying structure of the measure and recommend a cut-off point which might provide adequate sensitivity and specificity for screening purposes. Study Two aimed to explore the feasibility of developing an observational version of the GDQ for use in primary care settings. The use of existing data allowed for an initial exploration of the validity of the GDQ without exposing participants to invasive and potentially unnecessary medical procedures.  The inclusion of multiple data sets produced a reasonable sample size adequate for factor analysis, and allowed the inclusion of multiple genetic syndromes, reflective of the heterogenous nature of the ID population. The results of Study One revealed that there were significant differences in GDQ scores between children who had and had not experienced GORD recently. Study One also established a scoring strategy and clinical cut off for the GDQ which achieved a sensitivity score of .86. Study Two demonstrated that the application of the GDQ as a brief observational tool for a naïve observer does not correlate to parent/carer reported GDQ scores. Similarly, Study Two also demonstrated that observed coded behaviours do not differentiate between children who score above and below cut off on the GDQ. These findings suggest that brief observation by a naïve observer may not be sufficient to identify behaviours which are indicative of GORD in children with ID.

The first goal of Study One was to explore the factor structure of the GDQ, as this had not been undertaken previously. Five factors were identified using an exploratory factor analysis approach. Factor one was identified as the factor which contributed most to the model and included items relating to chewing, salivating, putting fingers in mouth, grinding teeth, scratching at the throat, crying and coughing. The selection of items included in factor one suggest that factor one scores may be indicative of distress and pain located in the mouth and throat. Factor two included the behaviours of lying on the stomach, arching of the back, and fidgeting. Based upon the inclusion of back arching and fidgeting, the co-occurance of which was referred to as “a non-verbal equivalent of heartburn” by Czinn and Blanchard (2013), one might conclude that factor two scores were indicative of heartburn. However, if factor two scores were heartburn related then a significant difference in factor two scores between children with and without a reported diagnosis of GORD would have been expected (Hassal, 2001; NICE, 2015). Given that such a difference was not found, it is plausible that factor two relates to either generalised pain related behaviours, stereotyped and repetitive behaviours, or hyperactivity, which are commonly reported in children with ID (Taanila, Ebeling, Heikura, & Järvelin, 2003). Factor three groups together children refusing food, appearing indecisive about food, and crying. The grouping of behaviours in factor three appears to be related to changes in appetite and meal time behaviour. Factor four groups together several known indicators of GORD; night waking, bad breath, and increased fluid intake. Based upon the grouping of these three behaviours, factor four may detect cases involving sleep problems such as sleep apnea. A review of the literature suggests that GORD and sleep apnea may exist in a mutually reinforcing relationship (Demeter & Pap, 2004). Factor five groups the remaining items; frequency of respiratory infection, which has a known association with GORD (Reyes, Cash, Green & Booth, 1993), and sleeping seated or propped up at night. Study one has identified a factor structure which fits with the current understanding of GORD. There may be potential that with further study the underlying factors of the GDQ could improve the usefulness of the tool and aid definition and delineation of atypical variants of GORD.

The second aim of Study One was to explore the validity of the GDQ as a tool for detecting GORD. The results of Study One demonstrated significant differences in both the total GDQ scores, and all factor scores except factor two, between children with and without a reported diagnosis of GORD. This was further supported by the ROC analysis which produced a significant area under the curve score for both proposed methods of scoring the GDQ. The findings of Study One establish that the GDQ is capable of distinguishing between children with and without GORD according to parental report of diagnosis. Although parental report of diagnosis is not equivalent to medical diagnostic procedures, the significant association between GDQ scores and parental report of diagnosis reported here offers an important indicator that the GDQ is a promising tool requiring further clinical validation. The findings of Study One suggest that the GDQ may be a potentially useful tool in supporting clinicians to identify children who would benefit from further investigation of GORD symptoms. It is important that GORD is detected and treated in people who cannot self-report their symptoms, given the established associations between untreated pain, behaviours that challenge, and reduced quality of life (Carr & DeSchryver, 2007; Wiggs & Stores, 1996; Beadle-Brown, Murphy & DiTerlizzi, 2009). A study comparing GDQ scores to GORD diagnosis, as established by gold standard medical procedures, is now required in order to support a recommendation for the GDQ as a tool to be used in clinical practice.

The results of Study One also established a clinical cut-off for the GDQ. The cut-off suggested by Youden's Index (YI) was > 8 which produced a sensitivity of .83 and a specificity of .56. This cut off should be applied to the amended total which omits items one, three and four. Similar sensitivity and specificity can also be achieved using a cut-off of > 12 for the full GDQ score. YI is a purely statistical strategy for determining cut-off choice and does not include any clinical judgement. A perfect test would produce both sensitivity and specificity values of 1, but the reality is that sensitivity is necessarily gained at the expense of specificity, or specificity at the expense of sensitivity (Watson & Petrie, 2010). For the purposes of a screening test, correct identification of children who may be displaying GORD symptomology (sensitivity) is more important than exclusion of children are not displaying such symptoms (specificity). Although one would not wish to expose children unnecessarily to the invasive health procedures required to confirm a GORD diagnosis, the GDQ is intended for use as a screening tool to be used alongside clinical judgement rather than being used as a diagnostic tool in isolation. The primary purpose of the GDQ is to support clinicians in recognising GORD symptoms in children who may otherwise go undiagnosed. As discussed in the introduction, the potential consequences of failing to detect GORD can be severe both in terms of behavioural consequences (Carr & DeSchryver, 2007; Allen, Lowe, Brophy & Moore, 2009; Beadle-Brown, Murphy & DiTerlizzi, 2009), and potential health consequences (Solaymani-Dodaran, Logan, West, Card & Coupland, 2004). Thus in the case of the

GDQ it is justifiable, and arguably preferable, to use a cut-off with sub-optimal specificity in order to maximise sensitivity.

The final aim of Study One was to explore the discriminant validity of the GDQ. The results of Study One found no association between GDQ scores and diabetes, liver problems, or heart problems as reported in the previous month. However there were significant differences in the rates of GORD, bowel problems, epilepsy, respiratory problems, and skin problems between children above and below GDQ cut offs. Importantly, the analysis of the strengths of the differences in scores utilising odds ratios, demonstrated that the odds ratio of a child above clinical cut off were higher for GORD than for any other assessed health problem, suggesting some discriminant validity for the GDQ. Many of the other health conditions which are associated with scores above clinical cut off on the GDQ have a known overlap with GORD, either because of shared symptoms or because they are commonly co-morbid. Bowel problems such as Irritable Bowel Syndrome (IBS) are known to be a common co-morbidity of GORD, occurring in nearly half of GORD patients (Kennedy et al., 1998; Frissora & Kick, 2005). Similarly, there is a body of literature evidencing the links between respiratory problems and GORD, with GORD thought to be a potential cause of respiratory infection in some cases.

Study One utilised historical data in the assessment of the GDQ. A key limitation is that no direct clinical data were collected, instead Study One is reliant on the accuracy of parental reports. Hence assumptions are made that parents have either received a clinical diagnosis of GORD for their child, or have identified recent symptoms being experienced by their child, many of whom are minimally verbal, and correctly identified those symptoms as being indicative of GORD, something which it has already been acknowledged, even clinicians sometimes struggle with. However, by making use of these historical data, Study One has achieved a large and diverse sample. Had the current study recruited a new sample specifically for the purposes of assessing the GDQ, it is unlikely that as many children would have been recruited as have been reported here. Study One has demonstrated a five factor structure underlying the GDQ, and has recommended a new scoring strategy, omitting items one, two and four which show no significant associations with recent GORD. Using the new scoring strategy a clinical cut off has been recommended which shows adequate sensitivity and specificity for using the GDQ as a screening tool to identify children who might benefit from further medical investigation of GORD. Further research is needed to assess the utility of the GDQ in medical settings, and to test the suggested cut off against clinical diagnosis of GORD as opposed to parent report, however the current study has provided adequate evidence to suggest that such research would be a worthwhile endeavour.

Study Two explored the feasibility of a brief observer version of the GDQ which could be used by a naïve observer such as a clinician. The development of a coding scheme based on behaviours from the GDQ, and establishing of acceptable kappa values between two naïve observers, suggest that many of the behaviours from the GDQ can be reliably observed, even by people who are unfamiliar with the child. Some of the items from the GDQ could not be developed into corresponding observational codes because they were not explicit behaviours, such as bad breath, or respiratory infections. Other items were unlikely to be observed given the nature of the footage being used, such as those items relating to behaviour around food or sleeping. Finally, three items were developed into codes, but were not observed in any of the analysed footage; swallowing, drinking, and gagging/regurgitating. The most commonly observed behaviours were fidget behaviours. Both fidgeting with the hands and fidgeting with the body, such as bouncing or swinging legs, were observed in at least half of the sample. The high frequency of fidgeting in Study Two gives some credence to the argument that fidgeting may generally be a high frequency behaviour in this population, which could mean that it might not be as useful an indicator of pain as it is in typically developing children (Czinn & Blanchard, 2013). Fidgeting or restless behaviours may only be indicative of pain in cases where they represent a deviation from the child's usual presentation, which would rely upon the knowledge of someone such as a caregiver who knows the child well.

The primary results of Study Two demonstrated that whilst many behaviours from the GDQ were observable, the scores acquired through observation were not associated with parent/caregiver GDQ scores. This finding may begin to offer some explanation as to why GORD might be under diagnosed in children who cannot self-report pain. It may be that brief observation might not be sufficient for detecting the behaviours which are most indicative of GORD, making it difficult to detect GORD related behaviours in a routine clinical appointment. However, there are limitations to the current study. Study two, similar to study one, was conducted using historical data, in order to achieve an adequate sample size, multiple experimental paradigms were included in the footage which was selected for analysis. Although attempts were made to limit the inclusion of experimentally induced stress behaviours in the analysed footage, this cannot be guaranteed. It is also of note that within the sample the group of children who scored above cut-off on the GDQ was relatively small. In conclusion, while study two might suggest that the use of the GDQ items to inform observation by a naïve observer does not yield the same results as parent report, the study is by no means robust enough to draw any larger conclusions regarding the usefulness of observation as a means of detecting GORD. The results of study one indicate that there are behavioural indicators of GORD which can be detected, it might be that parents are best placed to notice these behaviours, but that does not mean that clinicians could not be enabled to do so also, given the right tools. Further research to this end may still be of significant clinical benefit.

Health inequalities remain a key concern for those seeking to improve the lives of people with ID. Improved tools and training for clinicians are clearly required in order to improve recognition and diagnosis of treatable disorders in people with ID, and the studies reported here suggest that the GDQ may be one tool to improve recognition of gastric pain in this population. However, these studies also highlight the importance of parent or caregiver knowledge in the diagnostic process. Unfortunately, increasing the contribution that caregivers can make to the diagnostic process is not something that can be easily resolved. Parents and caregivers, when they are available, already frequently act as knowledgeable advocates for the people under their care, but the literature suggests that they may not always be consulted, or their reports may not be believed (Lewis, Gaffney & Wilson, 2017). Tools such as the GDQ may help to address this by formalising the caregiver report into a validated measure that a clinician can easily interpret. However addressing the broader issue of increasing the collaboration between health care professionals and care givers may require a much more systemic cultural shift.

## 2.6. References

Allen, D., Lowe, K., Brophy, S., & Moore, K. (2009). Predictors of restrictive reactive strategy use in people with challenging behaviour. Journal of Applied Research in Intellectual Disabilities, 22(2), 159-168.

Anders, P. L., & Davis, E. L. (2010). Oral health of patients with intellectual disabilities: a systematic review. *Special Care in Dentistry*, *30*(3), 110-117.

Andy, F. (2005). Discovering statistics using SPSS. *Saga Publication Ltd*, 521-570.

Arron, K., Oliver, C., Moss, J., Berg, K., & Burbidge, C. (2011). The prevalence and phenomenology of self-injurious and aggressive behaviour in genetic syndromes. *Journal of Intellectual Disability Research*, *55*(2), 109-120.

Beadle-Brown, J., Murphy, G., & DiTerlizzi, M. (2009). Quality of life for the Camberwell cohort. Journal of Applied Research in Intellectual Disabilities, 22(4), 380-390.

Böhmer, C. J. M., Klinkenberg-Knol, E. C., Niezen-de Boer, M. C., & Meuwissen, S. G. M. (2000). Gastroesophageal reflux disease in intellectually disabled individuals: how often, how serious, how manageable?. The American journal of gastroenterology, 95(8), 1868-1872.

Breau, L. M., McGrath, P. J., Camfield, C., Rosmus, C., & Finley, G. A. (2000). Preliminary validation of an observational checklist for persons with cognitive impairments and inability to communicate verbally. *Developmental medicine and child neurology*, *42*(9), 609-616.

Carr, E. G., & Owen-DeSchryver, J. S. (2007). Physical illness, pain, and problem behavior in minimally verbal people with developmental disabilities. Journal of Autism and Developmental Disorders, 37(3), 413-424.

Crosta, Q. R., Ward, T. M., Walker, A. J., & Peters, L. M. (2014). A review of pain measures for hospitalized children with cognitive impairment. *Journal for Specialists in Pediatric Nursing*, *19*(2), 109-118.

Czinn, S. J., & Blanchard, S. (2013). Gastroesophageal reflux disease in neonates and infants. *Pediatric Drugs*, *15*(1), 19-27.

Demeter, P., & Pap, A. (2004). The relationship between gastroesophageal reflux disease and obstructive sleep apnea. *Journal of gastroenterology*, *39*(9), 815-820.

Emerson, E., & Baines, S. (2011). Health inequalities and people with learning disabilities in the UK. *Tizard Learning Disability Review*, *16*(1), 42-48.

Findlay, L., Williams, A. D. C., & Scior, K. (2014). Exploring experiences and understandings of pain in adults with intellectual disabilities. *Journal of Intellectual Disability Research*, *58*(4), 358-367.

Fluss, R., Faraggi, D., & Reiser, B. (2005). Estimation of the Youden Index and its associated cutoff point. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *47*(4), 458-472.

Frissora, C. L., & Koch, K. L. (2005). Symptom overlap and comorbidity of irritable bowel syndrome with other conditions. *Current gastroenterology reports*, *7*(4), 264-271.

Glover, G., Williams, R., Heslop, P., Oyinlola, J., & Grey, J. (2017). Mortality in people with intellectual disabilities in England. *Journal of Intellectual Disability Research*, *61*(1), 62-74.

Gotham, K., Risi, S., Pickles, A., & Lord, C. (2006). The Autism Diagnostic Observation Schedule (ADOS). *Journal of Autism and Developmental Disorders*.

Hall, S. S., Arron, K., Sloneem, J., & Oliver, C. (2008). Health and sleep problems in Cornelia de Lange syndrome: a case control study. *Journal of Intellectual Disability Research*, *52*(5), 458-468.

Hassall, E. (2005). Decisions in diagnosing and managing chronic gastroesophageal reflux disease in children. The Journal of pediatrics, 146(3), S3-S12.

Haveman, M., Heller, T., Lee, L., Maaskant, M., Shooshtari, S., & Strydom, A. (2010). Major health risks in aging persons with intellectual disabilities: an overview of recent studies. *Journal of Policy and Practice in Intellectual Disabilities*, *7*(1), 59-69.

Hermann, C., Zohsel, K., Hohmeister, J., & Flor, H. (2008). Dimensions of pain-related parent behavior: development and psychometric evaluation of a new measure for children and their parents. *PAIN®*, *137*(3), 689-699.

Heslop, P., Blair, P., Fleming, P., Hoghton, M., Marriott, A., & Russ, L. (2013). Confidential Inquiry into premature deaths of people with learning disabilities (CIPOLD). *Bristol: Norah Fry Research Centre*.

Kennedy, T. M., Jones, R. H., Hungin, A. P. S., O'flanagan, H., & Kelly, P. (1998). Irritable bowel syndrome, gastro-oesophageal reflux, and bronchial hyper-responsiveness in the general population. *Gut*, *43*(6), 770-774.

Kushlick, A., Blunden, R., & Cox, G. (1973). A method of rating behaviour characteristies for use in large scale surveys of mental handicap. Psychological Medicine, 3(4), 466-478.

Lennox, N. G., Diggens, J. N., & Ugoni, A. M. (1997). The general practice care of people with intellectual disability: barriers and solutions. *Journal of Intellectual Disability Research*, *41*(5), 380-390.

Lewis, P., Gaffney, R. J., & Wilson, N. J. (2017). A narrative review of acute care nurses' experiences nursing patients with intellectual disability: Underprepared, communication barriers and ambiguity about the role of caregivers. *Journal of clinical nursing*, *26*(11-12), 1473-1484.

Liu, X. (2012). Classification accuracy and cut point selection. *Statistics in medicine*, *31*(23), 2676-2686.

Luzzani, S., Macchini, F., Valade, A., Milani, D., & Selicorni, A. (2003). Gastroesophageal reflux and Cornelia de Lange syndrome: typical and atypical symptoms. American Journal of Medical Genetics Part A, 119(3), 283-287.

Malviya, S., Voepel-Lewis, T., Merkel, S., & Tait, A. R. (2005). Difficult pain assessment and lack of clinician knowledge are ongoing barriers to effective pain management in children with cognitive impairment. *Acute Pain*, *7*(1), 27-32.

Martin, N., Oliver, C., & Hall, S. (1998). Obswin: Software for the collection and analysis of observational data. *Birmingham: University of Birmingham*.

Mencap (n.d.) *How common is learning disability?* Retrieved from https://www.mencap.org.uk/learning-disability-explained/research-and-statistics/how-common-learning-disability [accessed on 5 April 2019]

Merkel, S. I., Voepel-Lewis, T., Shayevitz, J. R., & Malviya, S. (1997). The FLACC: a behavioral scale for scoring postoperative pain in young children. *Pediatric nursing*, *23*(3), 293-298.

Morin, D., Mérineau-Côté, J., Ouellette-Kuntz, H., Tassé, M. J., & Kerr, M. (2012). A comparison of the prevalence of chronic disease among people with and without intellectual disability. *American Journal on Intellectual and Developmental Disabilities*, *117*(6), 455-463.

National Institute for Health and Care Excellence. (2015). *Gastro-oesophageal reflux disease in children and young people: diagnosis and management.* London: National Institute for Health and Care Excellence (UK).

NHS long term plan 2018. *The NHS Long Term Plan.* London: NHS England (UK)

Oliver, C. & Wilkie, L. (2005). *Gastro-Oesophageal Reflux Questionnaire*. University of Birmingham, Birmingham.

Oliver, C., Berg, K., Moss, J., Arron, K., & Burbidge, C. (2011). Delineation of behavioral phenotypes in genetic syndromes: characteristics of autism spectrum disorder, affect and hyperactivity. *Journal of Autism and Developmental Disorders*, *41*(8), 1019-1032.

Reyes, A. L., Cash, A. J., Green, S. H., & Booth, I. W. (1993). Gastrooesophageal reflux in children with cerebral palsy. *Child: care, health and development*, *19*(2), 109-118.

Richards, C., Oliver, C., Nelson, L., & Moss, J. (2012). Self-injurious behaviour in individuals with autism spectrum disorder and intellectual disability. *Journal of Intellectual Disability Research*, *56*(5), 476-489

Solaymani-Dodaran, M., Logan, R. F. A., West, J., Card, T., & Coupland, C. (2004). Risk of oesophageal cancer in Barrett's oesophagus and gastro-oesophageal reflux. *Gut*, *53*(8), 1070-1074.

Stinson, J. N., Kavanagh, T., Yamada, J., Gill, N., & Stevens, B. (2006). Systematic review of the psychometric properties, interpretability and feasibility of self-report pain intensity measures for use in clinical trials in children and adolescents. Pain, 125(1-2), 143-157

Taanila, A., Ebeling, H., Heikura, U., & Järvelin, M. R. (2003). Behavioural problems of 8-year-old children with and without intellectual disability. *Journal of pediatric neurology*, *1*(01), 015-024.

UN General Assembly, Convention on the Rights of Persons with Disabilities, 13 December 2006, A/RES/61/106, Annex I, retrieved from https://www.refworld.org/docid/4680cd212.html [accessed on 5 February 2019]

von Baeyer, C. L., Chambers, C. T., & Eakins, D. M. (2011). Development of a 10-item short form of the parents' postoperative pain measure: the PPPM-SF. *The Journal of Pain*, *12*(3), 401-406.

Watson, P. F., & Petrie, A. (2010). Method agreement analysis: a review of correct methodology. *Theriogenology*, *73*(9), 1167-1179.

Wiggs, L., & Stores, G. (1996). Severe sleep disturbance and daytime challenging behaviour in children with severe learning disabilities. Journal of Intellectual Disability Research, 40(6), 518-528.

# CHAPTER THREE, PUBLIC BRIEFING DOCUMENT:

# ASSESSING PAIN THROUGH OBSERVATION

## 3.1. Assessing pain through behavioural observation – why bother?

Everyone experiences pain differently, our experience of pain can be affected by our emotions, our expectations, and our previous experiences of pain or injury. Because of this, subjective ratings of pain are considered to be the “gold-standard” for pain assessment, even very young children will be asked by clinicians to provide their own ratings for their pain. However, not everyone can provide their own ratings, people might struggle to understand the rating system they are given, or they might have difficulty in communicating. But if we don’t have accurate ways to assess people’s pain without self-report then some people’s pain might be left untreated. Untreated pain can have serious negative consequences, it can affect our mood, our sleep, and our health.

When people aren’t able to self-report, usually someone else will attempt to judge how much pain they are in by looking at how they are behaving, for example, if someone is crying then we might think they are in more pain than if they are sitting quietly.

This document summarises research carried out by a Trainee Clinical Psychologist at the University of Birmingham that (i) explores and evaluates the tools researchers use when they are trying to observe pain, and (ii) assesses a new measure which is designed to help clinicians identify gastric pain.

## 3.2. Meta-analysis

### 3.2.1. What is a meta-analysis?

A meta-analysis is a way of bringing together all of the available data that has been collected on a certain topic. The findings are taken from each study that has researched the question, and put together into an analysis which tells us the average finding, taking into account things like how many people were recruited for each study, and the quality of each study.

### 3.2.2. What were you trying to find out?

The aims of the meta-analyses were:

1. To identify the most common observational assessments used in published studies that seek to quantify pain in children
2. To find out how well the most commonly used observational assessments compare to self-reports of pain.

### 3.2.3. What did you do?

Five databases were searched for words relating to “pain” “assessment” and “children”. All of the papers were reviewed to check they used observational measures to assess pain in children aged 1-18 years. Then all of the observational measures were recorded, along with how many studies used them.

Once all of the studies had been reviewed, papers were identified that reported a correlation between any of the five most common observational measures, and self-report measures of pain. The results of those correlations were put into a statistical programme to calculate a weighted average, a figure that attempts to summarise the overall findings of the literature.

### 3.2.4. What did you find out?

***What tools get used?*** There were 526 published studies that used observational measures to assess pain in children, and nearly a third of those studies used either the Children's Hospital East Ontario Pain Scale (CHEOPS; McGrath et al., 1985), the Visual Analogue Scale (VAS; Price, McGrath, Rafii & Buckingham, 1983), the Face, Legs, Activity, Crying and Consolability (FLACC; Voepel-Lewis, Shayevitz & Malviya, 1997), the Observer Pain Scale (OPS; Hannallah et al., 1987), or the Wong-Baker Faces Scale (WBF; Wong & Baker, 1988). Even though one of these five measures were used most of the time, there were still 62 different measures that were named, most of those measures were only used by one or two studies each.

***How well do observational pain measures correlate to self-report?*** The table below shows the results of the meta-analysis. The measure that correlated best to self-report was the WBF, but there were only four studies available that tested this, two of them had very low scores, and two of them had very high scores, so there was a lot of variability. This might be because the WBF was originally designed as a child-friendly tool for self-reporting pain, so it does not give the person using it any specific behaviours to look for. Because of that it is easy for people to interpret it in lots of different ways which might produce a lot of different results.

The FLACC scored slightly lower than the WBF with regards to correlation to self-report, however, there were more studies available which assessed this, and the finding was much more consistent across those studies. The FLACC is specifically designed as an observational assessment, and directs the user towards specific behaviours to score in order to produce the pain rating.

### 3.2.5. Conclusions

The Pediatric Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (PedIMMPACT) group only recommends seven different observational measures for use in clinical trials involving children (Von Baeyer & Spagrud, 2007), this is because, having reviewed the literature, they found that only seven measures had enough evidence to demonstrate that they were reliable and valid measures of pain. The FLACC and CHEOPS both feature in the PedIMMPACT recommendations, so it is positive that they are among the most commonly used measures. However, it also makes the number of measures identified in this review very concerning; if only seven measures have been recommended, why did this review find 62 being used?

## 3.3. Empirical Study

### 3.3.1. What were you trying to find out?

The aims of this study were:

1. To find out more about the Gastric Distress Questionnaire (GDQ) and how useful it could be as a screening measure for Gastro-oesophageal Reflux Disease (GORD)
2. To try to create a brief version of the GDQ that could be used by someone who does not know the child, for example, a nurse or doctor, to help them decide whether or not to refer the child for more specialist assessment

### 3.3.2. Who would that help?

GORD is a painful health condition, it happens when stomach acid keeps being brought back up into the oesophagus. Most people will experience reflux once in a while, particularly when they are very young, but when it keeps on happening over a much longer period of time it is called GORD, and this can be painful and can have much longer term health consequences. GORD seems to be much more common in people with ID than it is in the general population, but it also frequently goes undiagnosed and untreated, because of the difficulties some people with ID have in reporting their symptoms. If family members, carers, and clinicians had the tools to recognise when someone with ID was experiencing the pain associated with GORD, then hopefully that person would be more likely to get diagnosed and treated.

### 3.3.3. What did you do?

Firstly historical data from several previous studies was compiled. In all of the included studies, parents of children with ID had been asked about their child's recent health, including whether they had experienced GORD in the last month or not, and the parent had completed the GDQ. In total data from 599 children was collected. The results were then analysed to assess whether there was any difference in scores between children whose parents had reported having GORD recently, and those whose parents said they had not had GORD.

Factor analysis was carried out to look at how the different questions on the GDQ grouped together. An analysis of sensitivity and specificity was also carried out to help recommend a cut-off score that would make sure most children who did have GORD got picked up, without wrongly identifying too many of the children who did not have GORD.

To try to develop an observational tool, the questions from the GDQ were converted into detailed and specific descriptions of behaviours. Footage was then collected from previous studies where parents had completed the GDQ. For each child ten minutes of video footage was viewed, and researchers recorded every time a child did one of the behaviours, for example, putting their fingers or hands into their mouth. The researchers looked at how different behaviours they observed in each child, how many times they observed any behaviour from each child, and how many seconds out of the ten minutes the child spent doing any of the behaviours that were being recorded. These different behavioural scores were all compared to the GDQ score provided by the parents.

### 3.3.4. What did you find out?

Overall, higher scores on the GDQ were found in children with GORD than those without. The questions on the GDQ were found to group together into five groups, or factors. Factor two, which contained items asking if the child lies on their back, on their front, or fidgets a lot, was not found to relate to GORD, but all of the other factors were.

The best cut-off value identified was found by adding together all of the question scores except for the questions in factor two. When the total is calculated in this way, a score of nine or more correctly identified eight out of ten children with GORD.

The scores obtained through observation of the video footage did not show a good association with the GDQ scores.

### 3.3.5. What does this mean?

The GDQ could be a useful tool to help identify children who might benefit from medical assessment for GORD. However, because this study was based on parent reports of diagnosis, the findings are not enough to recommend that the GDQ be used in clinical practice. Further research is now needed to compare GDQ scores against actual medical assessments of GORD, but the current study does suggest that such research could be a worthwhile exercise.

# VOLUME 1 APPENDICES

## Appendix A: The Gastric Distress Questionnaire

**Instructions:**

- This questionnaire asks about behaviours sometimes shown by people with learning disabilities.
- Please read the questions and examples carefully and indicate how often each behaviour has occurred in the **last two weeks** by circling the appropriate answer.

| **Does the person you care for:** | **More than once an hour** | **Once an hour** | **Once a day** | **Once a week** | **Not occurred** |
|---|---|---|---|---|---|
| 1. Arch his/her back, lie over arms of chairs or people on his/her back?.......... | **4** | **3** | **2** | **1** | **0** |
| **2.** Lie over an object on his/her stomach? e.g. a side of an arm chair. ………………. | **4** | **3** | **2** | **1** | **0** |
| **3.** Salivate excessively? …………………………………… | **4** | **3** | **2** | **1** | **0** |
| **4.** Fidget, wriggle or move their body a great deal? ……………………… | **4** | **3** | **2** | **1** | **0** |
| **5.** Place their hands or fingers in back of their mouth? ………………… | **4** | **3** | **2** | **1** | **0** |
| **6.** Chew on his/her clothes, fingers, hands or other parts of the body, objects or material? ………………………………………………………………………… | **4** | **3** | **2** | **1** | **0** |
| **7.** Grind their teeth? ……………………………………………………………… | **4** | **3** | **2** | **1** | **0** |
| **8.** Scratch, hit, press or rub around the upper chest or throat? ……………………… | **4** | **3** | **2** | **1** | **0** |
| **9.** Drink, request or seek out an excessive amount of fluids? ……………………… | **4** | **3** | **2** | **1** | **0** |
| **10.** Cough, gag or regurgitate? …………………………………………………… | **4** | **3** | **2** | **1** | **0** |
| **11.** Appear in pain or discomfort (cry, groan or moan)? …………………………… | **4** | **3** | **2** | **1** | **0** |
| **12.** Refuse food even though they are probably hungry? ………………………… | **4** | **3** | **2** | **1** | **0** |

**13.** Does the person you care for appear indecisive about food (edging towards table or food then moving away repeatedly, taking food and putting it back)? *(please tick)*

**Yes** □ **No** □

**14.** Does the person you care for wake during the night?

**Never** **Once a week** **Most nights** **Every night**

**15.** Does the person you care for sleep sitting or propped up?

**Never** **Once a week** **Most nights** **Every night**

**16.** Does the person you care for seem to have bad breath?

**Never** **Once a week** **At the same time everyday** **All day every day**

**17.** Has the person you care for prone to respiratory tract infections? *(please tick)*
**Yes** □ **No** □

If 'yes' please indicate how often they occur:

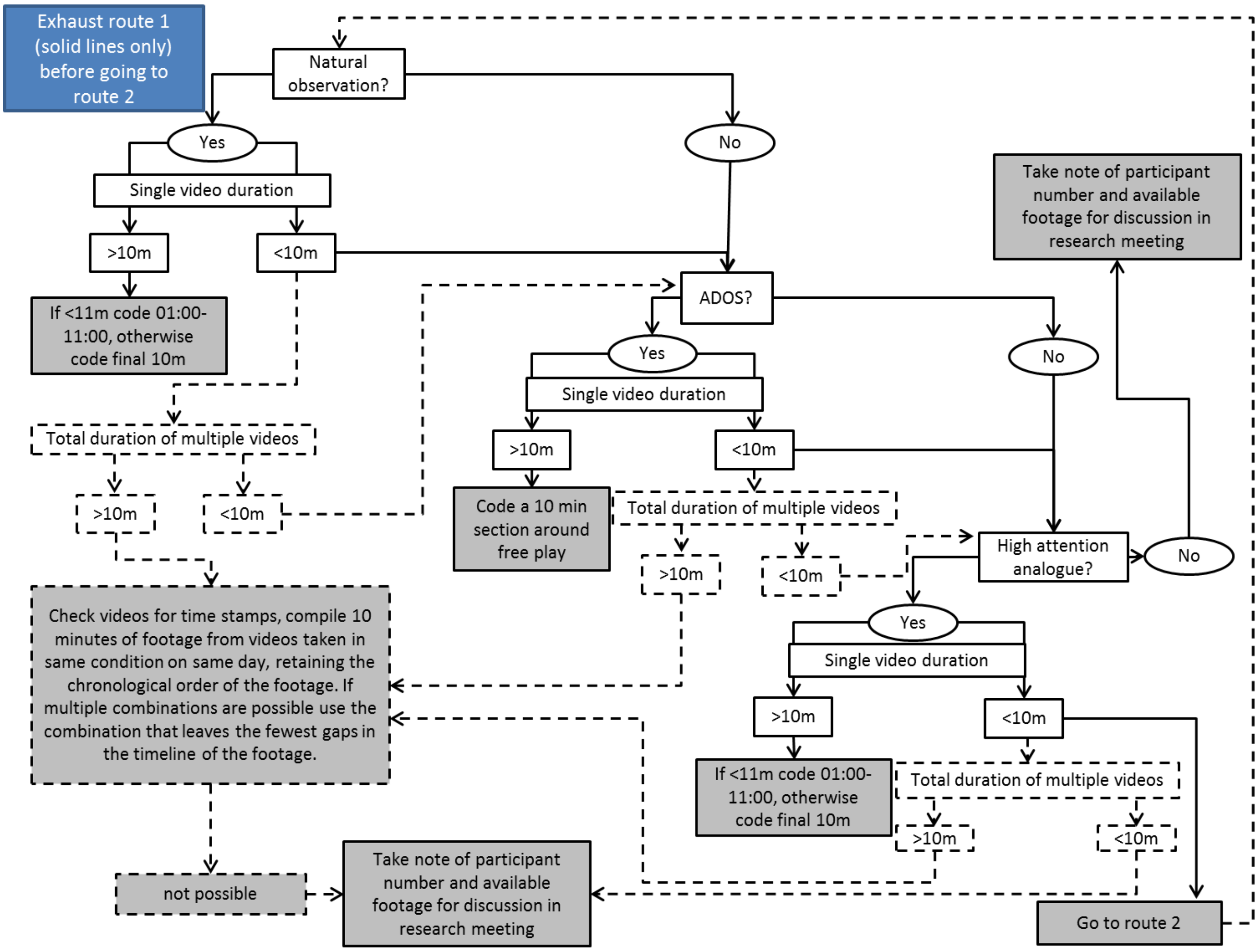
**Monthly** **Quarterly** **Every six months** **Annually**

Other (please specify)______________________________

## Appendix B: Protocol for extraction of correlation values for meta-analysis

| | |
|---|---|
| **Overlapping samples:** | In cases of multiple papers reporting the same sample then only one paper will be included. Selection will first be based on the paper reporting the most analysis of pain assessment will be selected, or if both papers report the same analyses, then the paper with the largest n will be chosen. |
| **Rater selection:** | In cases where correlation values are reported for multiple raters relating to the same child, clinician raters will be selected on the basis that the majority of obs scales are designed for clinical use. If multiple clinical raters are reported then most senior clinician will be selected. |
| **Self-report selection:** | If multiple self-report scales are reported separately then selection will be based on largest n first, or in the case of equal numbers then the scale which results in the best quality rating for the "self-report" domain will be usd |
| **Time selection:** | If multiple time points are reported then time point during painful procedure or time point that is closest to the completion of procedure will be selected |
| **Procedure selection:** | In cases where correlation values are reported for multiple painful procedures then both procedures will be reported separately if between group comparison, however for within group comparisons then control condition (condition which is described as replicating standard practice) will be selected to avoid inclusion of same child at multiple points in analysis |
| **Age selection:** | If multiple non-overlapping age groups are reported then each will be included separately in meta-analysis |

## Appendix C: Protocol for choosing video footage for empirical paper study two


Exhaust route 1 (solid lines only) before going to route 2
Natural observation?
Yes
No
Single video duration
>10m
<10m
If <11m code 01:00-11:00, otherwise code final 10m
Total duration of multiple videos
>10m
<10m
Check videos for time stamps, compile 10 minutes of footage from videos taken in same condition on same day, retaining the chronological order of the footage. If multiple combinations are possible use the combination that leaves the fewest gaps in the timeline of the footage.
not possible
Take note of participant number and available footage for discussion in research meeting
ADOS?
Yes
Single video duration
>10m
<10m
Code a 10 min section around free play
Total duration of multiple videos
>10m
<10m
No
Take note of participant number and available footage for discussion in research meeting
High attention analogue?
No
Yes
Single video duration
>10m
<10m
If <11m code 01:00-11:00, otherwise code final 10m
Total duration of multiple videos
>10m
<10m
Go to route 2

## Appendix D: Reviewed literature for meta-analysis with included measures and sample details

| Study | Informant | Measure 1 | Measure 2 | Measure 3 | Measure 4 | Measure 5 | Measure 6 | Self-report? | Sample size | Youngest participant (months) | Oldest participant (months) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Abdel-Ghaffar et al. (2011)** | Not named | CHEOPS | | | | | | no | 36 | 60 | 84 |
| **Abdelhalim et al. (2013)** | Researcher | CHEOPS | | | | | | no | 120 | 36 | 84 |
| **Abd-Elshafy et al. (2015 )** | Not named | OPS | | | | | | no | 50 | 48 | 144 |
| **Abdul et al. (2017)** | Researcher | SEM | | | | | | no | 50 | 48 | 96 |
| **Abdulhameed et al. (1989 )** | Clinician | VAS | | | | | | yes | 30 | 96 | 168 |
| **Acar et al. (2012 )** | Researcher | CHEOPS | | | | | | no | 50 | 24 | 120 |
| **Akhtar et al. (2014)** | Researcher | CHEOPS | | | | | | no | 60 | 60 | 144 |
| **Akin et al. (2010 )** | Not named | CHIPPS | | | | | | no | 60 | 24 | 96 |
| **Akinci et al. (2005)** | Clinician | BPS | | | | | | no | 22 | 36 | 192 |
| **Akkaya et al. (2009 )** | Clinician | OPS | CRS4 | | | | | no | 66 | 36 | 108 |
| **Akoglu et al. (2006 )** | Clinician | CHEOPS | | | | | | no | 46 | 24 | 144 |
| **Alhashemi et al. (2006 )** | Clinician | OPS | | | | | | no | 40 | 36 | 180 |
| **Alhashemi et al. (2007 )** | Not named | OPS | | | | | | no | 40 | 36 | 132 |
| **Ali et al. (2013 )** | Clinician | CHEOPS | | | | | | no | 120 | 24 | 72 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Al-Sadek et al. (2014)** | Not named | CHEOPS | OPS | no | 108 | 36 | 84 |
| **Alsadek et al. (2015)** | Not named | OPS | CHEOPS | no | 60 | 24 | 84 |
| **Alwugyan et al. (2007 )** | Parent | VAS | | yes | 281 | 72 | 144 |
| **Al-Zaben et al. (2012 )** | Not named | OPS | | no | 60 | 12 | 84 |
| **Aminabadi et al. (2008 )** | Clinician | SEM | | no | 78 | 48 | 60 |
| **Aminabadi et al. (2011 )** | Clinician | SEM | | yes | 80 | 72 | 84 |
| **An et al. (2017)** | Not named | CHIPPS | | no | 100 | 12 | 72 |
| **Anatol et al. (1997 )** | Researcher | CHEOPS | | no | 168 | 60 | 144 |
| **Andersen et al. (2015 )** | Clinician | COMFORT | | no | 45 | 12 | 36 |
| **Andrzejowski et al. (2002 )** | Clinician | CRS4 | | yes | 133 | 60 | 144 |
| **Anninger et al. (2007 )** | Researcher | MBPS | | no | 88 | 12 | 144 |
| **Anouar et al. (2016)** | Clinician | CHEOPS | | no | 40 | 12 | 48 |
| **Antony et al. (2016)** | Not named | OPS | | no | 50 | 12 | 96 |
| **Apan et al. (2010 )** | Researcher | OPS | | no | 110 | 36 | 192 |
| **Arts et al. (1994 )** | Researcher | GABR | | yes | 180 | 48 | 192 |
| **Asaad et al. (2011)** | Not named | CHIPPS | | no | 90 | 60 | 120 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Asadi et al. (2016)** | Not named | CHEOPS | | no | 98 | 36 | 144 |
| **Ashkenazi et al. (2005 )** | Clinician | CHEOPS | | no | 193 | 24 | 156 |
| **Ashkenazi et al. (2006 )** | Clinician | CHEOPS | | yes | 178 | 24 | 168 |
| **Atabek et al. (2015 )** | Clinician | MBPS | | yes | 50 | 96 | 144 |
| **Atef et al. (2008)** | Clinician | OPS | | no | 40 | 36 | 120 |
| **Ates et al. (1998 )** | Not named | MPOPS | VAS | no | 30 | 36 | 132 |
| **Aydin et al. (2016 )** | Parent & Clinician | WBF | | yes | 120 | 84 | 144 |
| **Babl et al. (2009 )** | Clinician | FLACC | VAS | no | 36 | 12 | 60 |
| **Babl et al. (2012)** | Clinician | FLACC | | no | 76 | 18 | 42 |
| **Badali et al. (2000)** | Parent | FACES | | yes | 23 | 60 | 144 |
| **Baghdadi (1999 )** | Not named | SEM | | yes | 28 | 72 | 144 |
| **Baghdadi (2000 )** | Clinician | SEM | | yes | 16 | 108 | 144 |
| **Bahorski et al. (2015 )** | Not named | CHEOPS | | yes | 173 | 18 | 204 |
| **Bai et al. (2004)** | Not named | VAS | OPS | no | 91 | 12 | 168 |
| **Bailey et al. (2015 )** | Parent | FLACC | | yes | 57 | 24 | 192 |
| **Balan et al. (2009 )** | Parent & Researcher | VAS | | yes | | 60 | 144 |
| **Barkan et al. (2014 )** | Parent | VAS | | no | 60 | 12 | 120 |
| **Bar-Meir et al. (2006)** | Clinician | FLACC | | no | 59 | 12 | 192 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Batra et al. (2003 )** | Researcher | OPS | | no | 120 | 24 | 96 |
| **Baxt et al. (2004 )** | Parent | FACES | CAS | no | 276 | 60 | 204 |
| **Baxter et al. (2011 )** | Not named | FLACC | | no | 51 | 12 | 36 |
| **Bayon-Mottu et al. (2014 )** | Parent & Clinician | FLACC | VAS | yes | 107 | 27 | 226 |
| **Bearden et al. (2012 )** | Parent & Clinician | VAS | | yes | 90 | 48 | 72 |
| **Benini et al. (2004 )** | Parent & Researcher | VAS | OWN | yes | 16 | 84 | 216 |
| **Benzon et al. (2015 )** | Clinician | FLACC | | no | 60 | 48 | 120 |
| **Beran et al. (2007)** | Parent, Clinician & Researcher | FPSR | | yes | 57 | 48 | 108 |
| **Berberich et al. (2009 )** | Parent & Clinician | FPSR | FLACC | yes | 41 | 48 | 72 |
| **Berde et al. (1991 )** | Researcher | CHEOPS | | yes | 35 | 36 | 84 |
| **Bergendahl et al. (2004 )** | Researcher | OPS | | no | 104 | 22.8 | 116.4 |
| **Beyaz et al. (2011 )** | Not named | FLACC | | no | 50 | 36 | 144 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Beyaz et al. (2011 )** | Not named | CHEOPS | | no | 120 | 36 | 180 |
| **Beyaz et al. (2012 )** | Clinician | CHEOPS | | yes | 100 | 36 | 180 |
| **Beyer et al. (1990 )** | Clinician | CHEOPS | | yes | 25 | 36 | 84 |
| **Bhatnagar et al. (2008 )** | Researcher | VAS | | no | 60 | 12 | 120 |
| **Bhattacharya et al. (2005 )** | Clinician | VAS | | no | 50 | 96 | 144 |
| **Birbicer et al. (2007 )** | Not named | CHEOPS | OPS | no | 60 | 24 | 120 |
| **Birnie et al. (2016 )** | Parent | FPSR | | yes | 171 | 96 | 144 |
| **Birnie, et al. (2017 )** | Parent | FPSR | | yes | 171 | 96 | 144 |
| **Bishai et al. (1999 )** | Parent & Clinician | VAS | | yes | 39 | 60 | 192 |
| **Bjorkman et al. (2012 )** | Researcher | FLACC | | yes | 29 | 60 | 180 |
| **Blanchais et al. (2017 )** | Clinician | CHEOPS | | no | 21 | 18 | 60 |
| **Boivin et al. (2008 )** | Parent & Clinician | CHEOPS | VAS | yes | 239 | 48 | 144 |
| **Bolton et al. (2002 )** | Clinician | CHEOPS | | yes | 30 | 17 | 72 |
| **Borazan et al. (2012 )** | Clinician | CRS4 | | no | 120 | 72 | 156 |
| **Borkar et al. (2005 )** | Clinician | OPS | | no | 50 | 36 | 156 |
| **Bosenberg et al. (2003 )** | Clinician | CRS4 | | yes | 110 | 46.8 | 144 |
| **Brahmbhatt et al. (2012)** | Parent & Clinician | NRS11 | | yes | 33 | 48 | 192 |
| **Breau et al. (2001 )** | Parent & Researcher | CFCS | VAS | yes | 123 | 52 | 80 |

|  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
| **Breau et al. (2002)** | Parent | NRS11 | NCCPC |  |  | no | 101 | 36 | 216 |
| **Breau et al. (2003 )** | Parent | NCCPC |  |  |  | no | 101 | 36 | 216 |
| **Breschan et al. (2005 )** | Clinician | CHIPPS |  |  |  | no | 182 | 12 | 84 |
| **Bridge et al. (2000)** | Not named | CHEOPS |  |  |  | yes | 30 | 48 | 144 |
| **Bringuier et al. (2009 )** | Parent & Researcher | CHEOPS | CHIPPS | FLACC | OPS | yes | 150 | 12 | 84 |
| **Brochard et al. (2009 )** | Not named | CHEOPS |  |  |  | yes | 34 | 24 | 180 |
| **Brudvik et al. (2017 )** | Parent & Clinician | NRS11 |  |  |  | yes | 243 | 36 | 180 |
| **Burns-Nader et al. (2016 )** | Researcher | FLACC |  |  |  | yes | 41 | 48 | 132 |
| **Burns-Nader et al. (2017 )** | Clinician | NRS6 |  |  |  | yes | 30 | 48 | 144 |
| **Burton et al. (1998 )** | Clinician | CHEOPS |  |  |  | no | 30 | 24 | 84 |
| **Caes et al. (2012 )** | Researcher | CFCS |  |  |  | no | 56 | 132 | 180 |
| **Cai et al. (2017 )** | Parent | PPPM |  |  |  | no | 204 | 12 | 72 |
| **Calis et al. (2014 )** | Not named | FLACC |  |  |  | no | 60 | 72 | 144 |
| **Canakci et al. (2017 )** | Not named | CHEOPS |  |  |  | no | 60 | 72 | 144 |
| **Canbulat et al. (2014 )** | Parent & Clinician | WBF |  |  |  | yes | 188 | 84 | 132 |
| **Cantekin et al. (2014 )** | Parent & Clinician | FLACC |  |  |  | no | 78 | 48 | 120 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Cao et al. (2011 )** | Not named | CHIPPS | | | | | | no | 59 | 72 | 96 |
| **Casey et al. (1990 )** | Researcher | OPS | | | | | | no | 60 | 24 | 120 |
| **Cassidy et al. (2001 )** | Parent, Clinician & Researcher | VAS | CHEOPS | CFCS | | | | yes | 161 | 48 | 72 |
| **Cassidy et al. (2002)** | Researcher | CHEOPS | CFCS | | | | | yes | 62 | 60 | 60 |
| **Celebi et al. (2013 )** | Clinician | OPS | | | | | | no | 60 | 36 | 72 |
| **Chadha et al. (2013 )** | Parent & Researcher | FLACC | WBF | | | | | yes | 69 | 36 | 144 |
| **Chalam et al. (2015)** | Not named | OPS | | | | | | no | 100 | 24 | 120 |
| **Chambers et al. (1996 )** | Parent | PPM | | | | | | yes | 110 | 84 | 144 |
| **Chambers et al. (1999 )** | Parent | PBCL | FACES | WBF | MFACE | KLPF | | yes | 75 | 60 | 144 |
| **Chambers et al. (2003)** | Parent | PPPM | | | | | | yes | 51 | 84 | 144 |
| **Chambers et al. (2005 )** | Parent & Clinician | PBCL | FACES | CAS | WBF | MFACE | KLPF | yes | 78 | 60 | 156 |
| **Chambers, et al. (1997)** | Parent | VAS | | | | | | no | 82 | 24 | 144 |
| **Chandler et al. (2013 )** | Researcher & Clinician | FLACC | | | | | | no | 112 | 24 | 72 |
| **Chang  (2005 )** | Parent | WBF | | | | | | yes | 101 | 24 | 192 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Chang (2008 )** | Parent | WBF | | | | | | yes | 69 | 24 | 192 |
| **Chang et al. (2015 )** | Clinician | CHEOPS | FLACC | CFCS | TPPS | RIPS | PEPPS | no | 44 | 13 | 74 |
| **Choi et al. (2003 )** | Clinician | CHEOPS | | | | | | no | 63 | 24 | 144 |
| **Choi et al. (2016 )** | Not named | MAPS | | | | | | no | 41 | 13 | 60 |
| **Choy et al. (1999 )** | Parent & Clinician | VAS/CAS? | | | | | | yes | 34 | 12 | 168 |
| **Christiano et al. (1998 )** | Parent | TPPS | | | | | | no | 74 | 12 | 64 |
| **Ciftci et al. (2014 )** | Not named | CHEOPS | | | | | | no | 52 | 12 | 204 |
| **Cobb et al. (2009 )** | Parent & Clinician | VAS | | | | | | yes | 89 | 48 | 144 |
| **Cohen et al. (1997 )** | Parent, Clinician & Researcher | VAS | FACES | | | | | yes | 62 | 48 | 72 |
| **Cohen et al. (2004 )** | Clinician | VAS | | | | | | yes | 39 | 105.96 | 132.96 |
| **Cohen et al. (2009 )** | Parent, Clinician & Researcher | VAS | OWN | | | | | yes | 57 | 48 | 72 |
| **Cole et al. (2009 )** | Clinician | FLACC | | | | | | no | 46 | 12 | 36 |
| **Cordoni et al. (2001 )** | Clinician | VAS | | | | | | yes | 57 | 48 | 144 |
| **Costa et al. (2011 )** | Researcher | FLACC | | | | | | no | 160 | 16 | 83 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Cregg et al. (1996 )** | Researcher | OPS | | no | 43 | 36 | 180 |
| **Dak-Albab et al. (2016 )** | Researcher | FLACC | | no | 30 | 96 | 144 |
| **Dal et al. (2007 )** | Clinician | OPS | | no | 90 | 24 | 144 |
| **Dalens et al. (2001 )** | Not named | OPS | CRS4 | no | 22 | 24 | 134.4 |
| **Davis et al. (2000 )** | Clinician | OPDS | | no | 206 | 24 | 144 |
| **De et al. (2004 )** | Clinician | CHIPPS | | no | 90 | 24 | 72 |
| **De et al. (2010)** | Clinician | FPSR | CAS | yes | 131 | 60 | 180 |
| **de et al. (2014 )** | Parent | PPPM | | yes | 174 | 48 | 120 |
| **De Gennaro et al. (2004 )** | Not named | CHEOPS | | yes | 10 | 48 | 204 |
| **Deepika et al. (2012 )** | Researcher | SEM | | yes | 60 | 72 | 144 |
| **Demiraran et al. (2006 )** | Clinician | WBF | | no | 75 | 12 | 72 |
| **Depue et al. (2013 )** | Parent, Clinician & Researcher | FLACC | VAS | no | 91 | 24 | 84 |
| **Dewhirst et al. (2014 )** | Clinician | FLACC | OPS | no | 99 | 12 | 92.4 |
| **Disma et al. (2009 )** | Researcher | CHEOPS | | no | 73 | 12 | 72 |
| **Duflo et al. (2004 )** | Not named | CHEOPS | VAS | no | 27 | 48 | 204 |
| **El et al. (2011)** | Clinician | FLACC | | no | 80 | 24 | 144 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Elbay et al. (2016 )** | Parent | VAS | | yes | 40 | 72 | 144 |
| **Eldeen et al. (2016)** | Not named | CHEOPS | | no | 40 | 24 | 60 |
| **El-Fattah et al. (2013 )** | Parent | PPPM | | yes | 135 | 60 | 144 |
| **Elhakim et al. (2003 )** | Researcher & Clinician | VAS | CHEOPS | yes | 110 | 48 | 120 |
| **El-Hamid et al. (2017 )** | Clinician | FLACC | | no | 86 | 36 | 84 |
| **Elsey et al. (2017)** | Clinician | OPS | | no | 17 | 24 | 216 |
| **El-Sharkawi et al. (2012 )** | Researcher | FLACC | | yes | 48 | 60 | 84 |
| **Eltzschig et al. (2002 )** | Not named | OPS | | no | 81 | 24 | 144 |
| **Emmott et al. (2016 )** | Researcher | FLACC | | yes | 120 | 36 | 72 |
| **Enyedi et al. (2017 )** | Researcher | CHEOPS | | no | 50 | 13 | 91 |
| **Ericsson et al. (2006 )** | Parent & Clinician | CRS7 | | yes | 92 | 60 | 180 |
| **Erol et al. (2008 )** | Clinician | CHIPPS | | no | 40 | 12 | 84 |
| **Ertugrul et al. (2006 )** | Researcher | TPPS | | no | 45 | 12 | 84 |
| **Evans et al. (1995 )** | Clinician | CHEOPS | | yes | 30 | 48 | 180 |
| **Faiz et al. (2013 )** | Not named | OPS | | no | 84 | 48 | 156 |
| **Fallah et al. (2016 )** | Not named | MBPS | | no | 70 | 18 | 18 |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| **Faraoni et al. (2010 )** | Clinician | OPS |  | no | 40 | 12 | 168 |
| **Farion et al. (2008)** | Parent & Clinician | CRS4 |  | yes | 80 | 72 | 144 |
| **Farrag et al. (2015 )** | Clinician | VAS |  | no | 40 | 36 | 120 |
| **Fatovich et al. (1999 )** | Parent & Clinician | VAS | OWN | no | 136 | 12 | 120 |
| **Fearon et al. (1996 )** | Researcher | DAL |  | no | 56 | 28 | 81 |
| **Feda et al. (2010 )** | Researcher | SEM |  | yes | 40 | 84 | 131 |
| **Fekih et al. (2013 )** | Not named | CHEOPS |  | no | 75 | 12 | 72 |
| **Fernandes et al. (2012 )** | Not named | FLACC |  | no | 80 | 12 | 120 |
| **Finley et al. (2003 )** | Parent | PPPM |  | yes | 75 | 84 | 144 |
| **Fishman et al. (2005 )** | Parent & Clinician | WBF |  | yes | 24 | 24 | 144 |
| **Foster et al. (2002 )** | Parent | VAS | TQPM | yes | 50 | 98.4 | 154.8 |
| **Franck et al. (2015 )** | Parent & Researcher | FLACC | FPSR | yes | 76 | 48 | 72 |
| **Furuya et al. (2009 )** | Parent & Clinician | FACES |  | no | 73 | 72 | 180 |
| **Galinkin et al. (2002 )** | Parent & Researcher | CHEOPS | VAS | yes | 22 | 84 | 192 |
| **Gamis et al. (1989 )** | Clinician | CHEOPS |  | no | 34 | 24 | 192 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Gazal et al. (2007 )** | Researcher | CHEOPS | | no | 201 | 24 | 144 |
| **Gedam et al. (2013 )** | Researcher & Clinician | FLACC | | no | 350 | 12 | 30 |
| **Georgopoulos et al. (2012 )** | Parent | TQPM | | yes | 124 | 48 | 144 |
| **Ghai et al. (2009 )** | Researcher | OPS | | no | 44 | 24 | 120 |
| **Ghosh et al. (2011 )** | Not named | OPS | | no | 90 | 12 | 60 |
| **Gilbert-MacLeod et al. (2000 )** | Researcher | DAL | | no | 60 | 24 | 72 |
| **Girotra et al. (1990 )** | Researcher | CRS3 | | no | 40 | 12 | 132 |
| **Girwalkar-Bagle et al. (2015 )** | Clinician | CHIPPS | | no | 60 | 24 | 60 |
| **Gomez et al. (2013 )** | Researcher & Clinician | FLACC | | no | 29 | 12 | 18 |
| **Goodenough et al. (1997 )** | Clinician | GBCL | | yes | 10 | 48 | 81 |
| **Goodenough et al. (1999 )** | Parent | VAS | | yes | 110 | 36 | 180 |
| **Goodenough et al. (2000)** | Parent & Researcher | FACES | GBCL | yes | 24 | 48 | 72 |
| **Goodenough, et al. (1997 )** | Researcher | FACES | | yes | 121 | 36 | 204 |
| **Goodenough, et al. (1998 )** | Researcher | FACES | GBCL | yes | 121 | 36 | 204 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Goodman et al. (2003 )** | Parent & Researcher | FACES | CFCS | | yes | 96 | 108 | 180 |
| **Goubert et al. (2009 )** | Parent & Researcher | NRS11 | CFCS | | yes | 53 | 110.04 | 180 |
| **Granry et al. (1997 )** | Clinician | MFACE | CRS4 | | yes | 88 | 72 | 144 |
| **Gunes et al. (2004 )** | Clinician | CHEOPS | | | no | 99 | 12 | 120 |
| **Gupta et al. (2014)** | Not named | FLACC | | | no | 70 | 12 | 84 |
| **Gurkan et al. (2017 )** | Clinician | CHEOPS | | | no | 75 | 16 | 72 |
| **Ha et al. (2013 )** | Parent & Researcher | VAS | PBCL | | yes | 84 | 36 | 120 |
| **Hadi et al. (2015 )** | Clinician | OPS | | | no | 92 | 36 | 84 |
| **Halperin et al. (2000 )** | Researcher | MBPS | | | no | | 12 | 15.6 |
| **Hamers et al. (1999 )** | Parent, Clinician & Researcher | CHEOPS | FLACC | VAS | yes | 83 | 36 | 144 |
| **Hamid et al. (2010 )** | Researcher | OWN | | | no | 40 | 42 | 96 |
| **Hartrick et al. (2002 )** | Clinician | TPPS | FLACC | COMFORT | no | 51 | 13.2 | 61.32 |
| **Hasani et al. (2011 )** | Clinician | OPS | | | no | 45 | 12 | 108 |
| **Hashizume et al. (2001 )** | Clinician | OPS | | | no | 60 | 12 | 60 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Hashizume et al. (2007 )** | Clinician | OPS | | no | 40 | 12 | 60 |
| **Hay et al. (2009 )** | Parent & Researcher | WBF | OWN | yes | 23 | 48 | 216 |
| **Hee et al. (2003 )** | Clinician | CHEOPS | | yes | 120 | 96 | 180 |
| **Helgadottir et al. (2014 )** | Parent | PPPM | | yes | 93 | 36 | 84 |
| **Hendrickson et al. (1990 )** | Parent & Clinician | VAS | | yes | 46 | 12 | 192 |
| **Hennrikus et al. (1995 )** | Clinician | CHEOPS | | yes | 97 | 48 | 204 |
| **Hesselgard et al. (2006 )** | Clinician | BOPS | | no | 26 | 32.4 | 88.8 |
| **Hesselgard et al. (2007)** | Clinician | BOPS | CHEOPS | no | 76 | 12 | 84 |
| **Hiller et al. (2006 )** | Researcher | MAUN | | no | 120 | 12 | 108 |
| **Hirschfeld et al. (2013 )** | Parent | VAS | | no | 2276 | 36 | 120 |
| **Hoeffe et al. (2017 )** | Researcher | FLACC | | yes | 90 | 48 | 216 |
| **Holthusen et al. (1994 )** | Researcher | CHEOPS | | no | 25 | 31.2 | 117.6 |
| **Honarmand et al. (2008 )** | Clinician | CHEOPS | | no | 75 | 36 | 144 |
| **Honarmand et al. (2013 )** | Clinician | CHEOPS | | no | 120 | 24 | 180 |
| **Hong et al. (2017 )** | Clinician | CHEOPS | | no | 62 | 36 | 84 |
| **Hopkins et al. (1988 )** | Researcher | OWN | VAS | no | 111 | 12 | 60 |
| **Hosey et al. (2006 )** | Clinician | CHEOPS | | no | 407 | 27.6 | 177.6 |
| **Hosten et al. (2011 )** | Researcher | CHEOPS | | no | 70 | 12 | 72 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Hua et al. (2015 )** | Parent & Clinician | FLACC | VAS | no | 65 | 48 | 192 |
| **Huet et al. (2011 )** | Not named | OPS | | yes | 30 | 60 | 144 |
| **Huh et al. (2017 )** | Clinician | CHEOPS | FLACC | no | 75 | 36 | 120 |
| **Hullett et al. (2009 )** | Not named | FLACC | | no | 51 | 12 | 36 |
| **Hunt et al. (2004 )** | Parent & Clinician | PPP | | no | 140 | 12 | 216 |
| **Inal et al. (2012 )** | Parent & Clinician | FPSR | | yes | 120 | 72 | 144 |
| **Inal et al. (2012)** | Parent & Clinician | FPSR | | yes | 123 | 72 | 144 |
| **Inanoglu et al. (2009 )** | Clinician | CHEOPS | | no | 90 | 24 | 144 |
| **Ipp et al. (2004 )** | Parent, Clinician & Researcher | VAS | MPBS | no | 49 | 12 | 12 |
| **Ipp et al. (2006 )** | Parent & Clinician | VAS | | yes | 60 | 48 | 72 |
| **Isaac et al. (2006 )** | Researcher | CHEOPS | | no | 14 | 12 | 96 |
| **Ivani et al. (1996 )** | Clinician | OPS | | no | 42 | 12 | 120 |
| **Ivani et al. (2000 )** | Researcher | OPS | | no | 40 | 12 | 84 |
| **Ivani et al. (2002 )** | Clinician | OPS | | no | 40 | 12 | 84 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Ivani et al. (2003 )** | Clinician | CHIPPS | | | no | 60 | 12 | 84 |
| **Ivani et al. (2005 )** | Not named | CHIPPS | | | no | 60 | 12 | 84 |
| **Jaaniste et al. (2007 )** | Researcher | CFCS | | | yes | 78 | 84 | 144 |
| **Jagannathan et al. (2009 )** | Researcher | CHIPPS | | | no | 48 | 12 | 72 |
| **Jamali et al. (1994 )** | Clinician | OPS | | | yes | 45 | 12 | 84 |
| **James et al. (2017)** | Clinician | RCEM | | | yes | 91 | 96 | 192 |
| **Jensen et al. (2012 )** | Parent | VAS | | | yes | 100 | 33.6 | 153.6 |
| **Jeong et al. (2012 )** | Not named | CHEOPS | | | no | 60 | 24 | 96 |
| **Jongudomkarn et al. (2008 )** | Parent & Clinician | WBF | NRS11 | KKU | yes | 150 | 72 | 144 |
| **Joyce et al. (1994 )** | Researcher | PRS | NAPI | POPS | no | 65 | 12 | 36 |
| **Jylli et al. (1995 )** | Parent & Clinician | VAS | | | yes | 129 | 36 | 192 |
| **Kamath et al. (2013 )** | Not named | TPPS | | | no | 56 | 48 | 60 |
| **Kankkunen et al. (2003 )** | Parent | VAS | PPPM | | no | 315 | 12 | 72 |
| **Kankkunen et al. (2003)** | Parent | PPPM | VAS | | no | 315 | 12 | 72 |
| **Kankkunen et al. (2009 )** | Parent | PPPM | CRS5 | | no | 50 | 12 | 35 |
| **Kannojia et al. (2017 )** | Not named | CHEOPS | | | no | 90 | 24 | 84 |
| **Karaaslan et al. (2008 )** | Clinician | CHEOPS | | | no | 75 | 36 | 144 |
| **Karakoyunlu et al. (2015 )** | Clinician | OPS | | | no | 60 | 24 | 132 |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| **Karamese et al. (2014 )** | Not named | CHEOPS |  | no | 7 | 12 | 60 |
| **Kawaraguchi et al. (2006 )** | Clinician | CHEOPS |  | no | 36 | 36 | 84 |
| **Kaya et al. (2012 )** | Not named | CHIPPS |  | no | 60 | 12 | 120 |
| **Kazak et al. (2010 )** | Clinician | FLACC | OPS | no | 60 | 24 | 72 |
| **Keidan et al. (2003)** | Clinician | NRS11 |  | yes | 31 | 36 | 180 |
| **Keidan et al. (2005 )** | Researcher | FLACC |  | no | 47 | 36 | 180 |
| **Keller et al. (2006 )** | Researcher | UWCH |  | no | 100 | 12 | 60 |
| **Kelly et al. (2002 )** | Parent | VAS |  | yes | 78 | 96 | 180 |
| **Kelly et al. (2015 )** | Parent & Researcher | OPS |  | yes | 91 | 12 | 120 |
| **Kennedy et al. (1998 )** | Parent | VAS |  | no | 260 | 60 | 180 |
| **Khan et al. (2008 )** | Not named | TPPS |  | no | 60 | 13 | 53 |
| **Khosravi et al. (2006 )** | Parent & Clinician | CHEOPS | CRS5 | no | 60 | 24 | 84 |
| **Kim et al. (2003 )** | Clinician | CHEOPS |  | no | 51 | 24 | 84 |
| **Kim et al. (2011 )** | Clinician | CHEOPS |  | no | 64 | 36 | 84 |
| **Kim et al. (2012 )** | Parent & Researcher | VAS | WBF | yes | 42 | 60 | 120 |
| **Kim et al. (2014 )** | Parent & Clinician | FLACC | PPPM | no | 80 | 24 | 72 |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| **Kim et al. (2017)** | Researcher | FLACC |  | no | 44 | 12 | 84 |
| **Klein et al. (2002 )** | Clinician | CHEOPS |  | no | 51 | 24 | 96 |
| **Knutsson et al. (2006 )** | Parent & Clinician | VAS |  | yes | 100 | 36 | 108 |
| **Knutsson et al. (2006 )** | Parent & Clinician | VAS | CHEOPS | no | 295 | 18 | 24 |
| **Kocum et al. (2013 )** | Researcher | CHEOPS |  | no | 120 | 36 | 72 |
| **Koinig et al. (1999 )** | Clinician | GOPS |  | no | 56 | 18 | 84 |
| **Kokki et al. (1999 )** | Clinician | MAUN |  | yes | 59 | 12 | 144 |
| **Kokki et al. (2003)** | Parent | VAS | PPPM | no | 85 | 12 | 72 |
| **Kokki et al. (2004)** | Clinician | MAUN |  | yes | 56 | 36 | 180 |
| **Kokki et al. (2006 )** | Not named | MAUN |  | no | 8 | 13 | 153 |
| **Koner et al. (2011 )** | Clinician | CHIPPS |  | no | 84 | 12 | 84 |
| **Kotzer et al. (1998 )** | Researcher | CPS |  | yes | 93 | 96 | 216 |
| **Kreider et al. (2001 )** | Clinician | SEM |  | yes | 32 | 72 | 180 |
| **Kundra et al. (2006 )** | Clinician | APDS |  | no | 132 | 24 | 144 |
| **Lal et al. (2001 )** | Parent & Clinician | PMH |  | yes | 27 | 48 | 96 |
| **LaMontagne et al. (1991 )** | Clinician | VAS |  | yes | 13 | 96 | 216 |
| **Lassaletta et al. (1997)** | Clinician | OWN |  | no | 120 | 24 | 168 |
| **Ledowski et al. (2017)** | Clinician | FLACC |  | no | 31 | 24 | 48 |
| **Lee et al. (1996 )** | Not named | CHEOPS | PBCL | yes | 137 | 36 | 84 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Lee et al. (2010)** | Clinician | CHEOPS | | no | 93 | 24 | 168 |
| **Lee et al. (2012 )** | Not named | CHEOPS | | no | 32 | 36 | 120 |
| **Lee-Jayaram et al. (2010 )** | Parent | VAS | | no | 23 | 60 | 216 |
| **Lejus et al. (2001)** | Clinician | KRANE | | no | 261 | 24 | 0 |
| **Leong et al. (2007)** | Researcher | TPPS | OPS | no | 54 | 24 | 72 |
| **Li et al. (2016 )** | Clinician | CHEOPS | | no | 60 | 36 | 84 |
| **Li et al. (2017 )** | Clinician | CHEOPS | | no | 80 | 36 | 84 |
| **Liang et al. (2014)** | Not named | CHIPPS | | no | 90 | 36 | 84 |
| **Lilley et al. (1997 )** | Researcher | NFCS | BFACS | no | 15 | 18 | 18 |
| **Lin et al. (2009 )** | Researcher | CHEOPS | | no | 60 | 12 | 72 |
| **Louw et al. (2016 )** | Researcher | CHEOPS | | no | 16 | 60 | 152 |
| **Lullmann et al. (2010 )** | Parent & Clinician | VAS | | yes | 87 | 24 | 216 |
| **Lundeberg et al. (2006 )** | Researcher | CHEOPS | | no | 40 | 12 | 24 |
| **Luo et al. (2017 )** | Clinician | CHIPPS | | no | 93 | 12 | 60 |
| **Maciocia et al. (2003 )** | Parent & Clinician | VAS | WBF | yes | 73 | 48 | 168 |
| **Magaret et al. (2002)** | Parent | WBF | | yes | 101 | 60 | 204 |
| **Mahajan et al. (2004 )** | Researcher | OPS | | no | 80 | 24 | 96 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Malmgren et al. (2004 )** | Not named | OPDS | | | | no | 82 | 24 | 72 |
| **Marseglia et al. (2015)** | Not named | FLACC | | | | no | ? | 12 | 36 |
| **Maryam et al. (2017)** | Clinician | CHEOPS | | | | no | 80 | 60 | 144 |
| **Massaro et al. (2014 )** | Parent & Clinician | CHEOPS | NCCPC | DESS | | no | 40 | 36 | 216 |
| **Mattila et al. (2016 )** | Clinician | OPS | | | | no | 49 | 15.6 | 116.4 |
| **Maunuksela et al. (1986 )** | Clinician | CRS4 | | | | yes | 60 | 48 | 120 |
| **Maunuksela et al. (1987 )** | Parent | MAUN | MFACE | | | yes | 141 | 19.2 | 211.2 |
| **Maunuksela et al. (1988 )** | Clinician | MAUN | | | | yes | 100 | 12 | 192 |
| **Maunuksela et al. (1992 )** | Clinician | MAUN | | | | no | 90 | 36 | 144 |
| **Maunuksela et al. (1992 )** | Clinician | MAUN | | | | yes | 128 | 48 | 144 |
| **McCarthy et al. (2000 )** | Parent & Clinician | TPPS | VAS | NRS10 | | no | 100 | 12 | 60 |
| **McCarty et al. (2000 )** | Clinician | CHEOPS | | | | no | 114 | 12 | 130 |
| **McIntyre et al. (2012 )** | Parent | NRS11 | | | | no | 178 | 24 | 96 |
| **McJunkins et al. (2010 )** | Parent & Clinician | VAS | MOPS | CHEOPS | NCCPC | no | 11 | 36 | 144 |
| **McWilliams et al. (2007 )** | Clinician | CHEOPS | | | | no | 74 | 24 | 72 |
| **Memis et al. (2003 )** | Clinician | TPPS | | | | no | 45 | 12 | 60 |
| **Mikawa et al. (1996 )** | Researcher | OPS | | | | no | 90 | 60 | 144 |
| **Mikawa et al. (1997 )** | Researcher | OPS | | | | no | 90 | 24 | 132 |
| **Miller et al. (2010 )** | Parent & Clinician | FLACC | VAS | | | yes | 80 | 36 | 120 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Min et al. (2012 )** | Clinician | CHEOPS | | no | 44 | 36 | 156 |
| **Miner et al. (2007 )** | Clinician | CHEOPS | | no | | 18 | 72 |
| **Minute et al. (2012 )** | Researcher | FLACC | | yes | 97 | 48 | 120 |
| **Mitrakul et al. (2015 )** | Clinician | FLACC | | yes | 42 | 60 | 96 |
| **Mittal et al. (2015 )** | Clinician | MBPS | | yes | 102 | 60 | 144 |
| **Moadad et al. (2016 )** | Parent & Clinician | WBF | | yes | 48 | 48 | 144 |
| **Modaresi et al. (1996 )** | Researcher | IZF | | yes | 30 | 93.6 | 188.4 |
| **Mohamed (2015)** | Not named | CHEOPS | | no | 48 | 18 | 36 |
| **Mohan et al. (2015 )** | Clinician | FLACC | WBF | yes | 123 | 48 | 180 |
| **Mohebbi et al. (2014 )** | Parent | VAS | | no | 80 | 60 | 180 |
| **Moir et al. (2000 )** | Parent | WBF | | yes | 51 | 36 | 144 |
| **Moon et al. (2008 )** | Parent | FACES | | yes | 73 | 48 | 144 |
| **Morgan et al. (2001 )** | Parent & Clinician | OWN | | no | 42 | 12 | 58.8 |
| **Morteza et al. (2012)** | Not named | CHEOPS | | no | 70 | 60 | 180 |
| **Mott et al. (2008 )** | Parent | VAS | | yes | 42 | 36 | 168 |
| **Movahedi et al. (2007 )** | Not named | CHEOPS | | yes | 80 | 72 | 144 |
| **Munro et al. (1994 )** | Clinician | OWN | | no | 42 | 24 | 108 |
| **Murray et al. (1987)** | Clinician | OWN | | no | 40 | 48 | 156 |
| **Nader et al. (2004 )** | Researcher | FACES | CFCS | no | 43 | 37.2 | 94.2 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Naja et al. (2013 )** | Clinician | OPS | | no | 80 | 24 | 60 |
| **Nicodemus et al. (1991 )** | Clinician | OWN | | yes | 60 | 24 | 156 |
| **Nilsson et al. (2008 )** | Parent | FLACC | | yes | 80 | 60 | 192 |
| **Nilsson et al. (2009 )** | Clinician | FLACC | | yes | 80 | 84 | 192 |
| **Nishina et al. (2000 )** | Researcher | OPS | | no | 125 | 24 | 144 |
| **Noel et al. (2015 )** | Parent | NRS11 | | no | 49 | 120 | 216 |
| **Noel et al. (2017 )** | Parent | NRS11 | | no | 66 | 120 | 216 |
| **Noguchi(2006 )** | Researcher | FACES | | yes | 64 | 48 | 78 |
| **Norambuena et al. (2013 )** | Researcher | CHEOPS | | no | 60 | 12 | 60 |
| **Numanoglu et al. (2014 )** | Clinician | OPS | | yes | 52 | 24 | 84 |
| **Nyman et al. (2005 )** | Clinician | CRS4 | | no | 83 | 24 | 216 |
| **O'Brien et al. (2004 )** | Researcher | MBPS | | no | 120 | 12 | 12 |
| **Odabas et al. (2012 )** | Researcher | MBPS | | yes | 50 | 84 | 156 |
| **O'Flaherty et al. (2003 )** | Clinician | OPS | | no | 80 | 36 | 144 |
| **Ohashi et al. (2016 )** | Clinician | BOPS | | no | 40 | 12 | 72 |
| **Oksuz et al. (2017 )** | Clinician | FLACC | | no | 53 | 12 | 84 |
| **Olanipekun et al. (2015 )** | Parent | FLACC | VAS | no | 62 | 12 | 84 |
| **Ozbek et al. (2002 )** | Clinician | CHEOPS | | no | 109 | 12 | 108 |
| **Ozyuvaci et al. (2004 )** | Researcher | CHEOPS | | no | 60 | 36 | 144 |
| **Pan et al. (2005 )** | Clinician | VPS | | no | 100 | 60 | 120 |
| **Parameswari et al. (2010 )** | Clinician | FLACC | | no | 100 | 12 | 36 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Park et al. (2004 )** | Clinician | CRS4 | | no | 130 | 24 | 144 |
| **Paschos et al. (2006 )** | Clinician | CHEOPS | SEM | yes | 104 | 60 | 144 |
| **Passariello et al. (2004 )** | Clinician | CHEOPS | | no | 44 | 12 | 66 |
| **Paut et al. (2001 )** | Researcher | OPS | | yes | 40 | 72 | 132 |
| **Peden et al. (2003 )** | Clinician | TPPS | DPT | no | 40 | 12 | 60 |
| **Peden et al. (2005 )** | Clinician | DPT | | yes | 64 | 72 | 144 |
| **Pickford et al. (2000)** | Clinician | CRS4 | | no | 69 | 36 | 120 |
| **Pierce et al. (1997 )** | Clinician | CHEOPS | | no | 35 | 36 | 216 |
| **Pieters et al. (2010 )** | Parent & Clinician | CHEOPS | NRS11 | no | 42 | 36 | 84 |
| **Potts et al. (2017)** | Not named | FLACC | | yes | 224 | 48 | 216 |
| **Pour et al. (2017 )** | Clinician | FLACC | | no | 120 | 72 | 144 |
| **Primosch et al. (2001 )** | Clinician | CHEOPS | | yes | 40 | 89 | 191 |
| **Prosser et al. (1997 )** | Clinician | TPPS | | no | 90 | 13 | 53 |
| **Purday et al. (1996 )** | Researcher | OPS | | no | 120 | 24 | 120 |
| **Rabbitts et al. (2015 )** | Parent | NRS11 | | no | 915 | 24 | 216 |
| **Ragg et al. (2017 )** | Parent | NRS11 | | yes | 100 | 72 | 216 |
| **Rai et al. (2014 )** | Not named | SEM | | yes | 60 | 72 | 168 |
| **Rajasagaram et al. (2009)** | Parent & Clinician | NRS11 | | yes | 86 | 36 | 180 |
| **Ram et al. (2003 )** | Clinician | MBPS | | no | 102 | 36 | 120 |
| **Ram et al. (2006 )** | Clinician | MBPS | | yes | 62 | 60 | 156 |
| **Ramirez et al. (2015 )** | Clinician | CHEOPS | | no | 69 | 24 | 84 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Ramirez-Carrasco et al. (2017 )** | Clinician | FLACC | | no | 40 | 60 | 108 |
| **Rattaz et al. (2013 )** | Researcher | NCCPC | CFCS | no | 67 | 36 | 96 |
| **Redmann et al. (2017 )** | Clinician | FLACC | | no | 125 | 36 | 144 |
| **Reid et al. (1987 )** | Clinician | VAS | | no | 49 | 12 | 84 |
| **Reid, et al. (1995)** | Parent | VAS | | no | 176 | 24 | 144 |
| **Rice et al. (1990 )** | Researcher | OPS | | no | 40 | 18 | 132 |
| **Riddell et al. (2004 )** | Parent | VAS | | no | 49 | 12 | 18 |
| **Rieger et al. (1996 )** | Researcher | OPS | | no | 41 | 24 | 120 |
| **Risaw et al. (2017)** | Parent & Clinician | FLACC | WBF | yes | 210 | 48 | 72 |
| **Ritterman et al. (2014 )** | Not named | FLACC | | no | 7 | 22 | 36 |
| **Rizk et al. (2014)** | Not named | OPS | | no | 90 | 36 | 72 |
| **Rocha et al. (2003 )** | Researcher | FACS | | no | 163 | 56 | 68 |
| **Romsing, et al. (1996 )** | Parent | VAS | | yes | 100 | 36 | 180 |
| **Ronnerfalt et al. (1998 )** | Clinician | CRS4 | | yes | 29 | 48 | 108 |
| **Rosales et al. (2016 )** | Parent | PPPM | | no | 161 | 24 | 180 |
| **Rose et al. (1999 )** | Parent & Clinician | VAS | CHEOPS | yes | 57 | 72 | 144 |
| **Rubinstein et al. (2016 )** | Parent & Clinician | VAS | VAS | no | 68 | 12 | 120 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Sadhasivam et al. (2014 )** | Clinician | FLACC | yes | 149 | 72 | 180 |
| **Sakellaris et al. (2004 )** | Researcher | OPS | no | 45 | 72 | 120 |
| **Salgado et al. (2013 )** | Clinician | VAS | no | 41 | 24 | 156 |
| **Sammons et al. (2007 )** | Researcher | TPPS | yes | 86 | 24 | 60 |
| **Sandeep et al. (2016 )** | Researcher | SEM | yes | 100 | 84 | 156 |
| **Sanders et al. (2007 )** | Parent & Clinician | VAS | no | 53 | 18 | 48 |
| **Sargin et al. (2015 )** | Not named | NCCPC | no | 40 | 72 | 192 |
| **Sato et al. (2010 )** | Researcher | CHIPPS | no | 81 | 12 | 108 |
| **Saxena et al. (2014 )** | Not named | FLACC | no | 70 | 12 | 84 |
| **Saylan et al. (2014)** | Not named | CHIPPS | no | 40 | 24 | 120 |
| **Schmitz et al. (2015 )** | Parent | VAS | yes | 535 | 36 | 216 |
| **Schneider, et al. (1992 )** | Parent & Clinician | OUCHER | yes | 42 | 35 | 78 |
| **Schultz et al. (1999 )** | Clinician | PEPPS | no | 40 | 12 | 24 |
| **Schutzman et al. (1996 )** | Researcher | CHEOPS | no | 39 | 36 | 96 |
| **Sen et al. (2014)** | Not named | CHIPPS | no | 60 | 18 | 84 |
| **Senel et al. (2001 )** | Clinician | CRS3 | no | 60 | 12 | 84 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Sermet et al. (2016 )** | Not named | FLACC | | | | yes | 60 | 72 | 144 |
| **Sethna et al. (2005 )** | Clinician | CRS4 | | | | yes | 64 | 36 | 204 |
| **Sezen et al. (2014)** | Clinician | CHIPPS | | | | no | 68 | 24 | 84 |
| **Shaikh et al. (2015 )** | Not named | FLACC | | | | no | 60 | 12 | 144 |
| **Shamim et al. (2015 )** | Parent & Clinician | CHEOPS | TPPS | FLACC | OPS | no | 25 | 36 | 84 |
| **Shavit et al. (2008 )** | Clinician | AHPS | | | | yes | 75 | 36 | 180 |
| **Shehab et al. (2015 )** | Researcher | SEM | | | | yes | 100 | 108 | 144 |
| **Shi et al. (2017 )** | Not named | FLACC | | | | no | 178 | 36 | 144 |
| **Shirazi et al. (2016 )** | Clinician | FLACC | | | | no | 42 | 12 | 72 |
| **Siddiqui et al. (2013 )** | Clinician | CHEOPS | | | | no | 75 | 60 | 144 |
| **Sikka et al. (2015 )** | Parent & Clinician | NRS11 | | | | yes | 50 | 60 | 216 |
| **Sikorova et al. (2011 )** | Clinician | CHEOPS | | | | yes | 60 | 60 | 120 |
| **Singer et al. (2002)** | Parent & Clinician | VAS | | | | yes | 63 | 48 | 84 |
| **Singh et al. (2012 )** | Clinician | FLACC | | | | no | 90 | 12 | 120 |
| **Singh et al. (2012 )** | Clinician | FLACC | | | | no | 80 | 12 | 120 |
| **Sinha et al. (2006 )** | Parent | VAS | | | | yes | 240 | 72 | 216 |
| **Sinha et al. (2009 )** | Not named | OPS | | | | no | 96 | 36 | 144 |
| **Sixou et al. (2009 )** | Clinician | VAS | | | | yes | 50 | 72 | 192 |
| **Smith et al. (1979 )** | Clinician | CRS5 | | | | yes | 212 | 180 | 204 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Smith et al. (1997 )** | Parent, Clinician & Researcher | VAS | | | | | | yes | 240 | 12 | 216 |
| **Smith et al. (1998 )** | Parent, Clinician & Researcher | VAS | | | | | | yes | 40 | 12 | 204 |
| **Smith et al. (2009)** | Clinician | FLACC | | | | | | no | 178 | 84 | 144 |
| **Soetenga et al. (1999 )** | Parent & Clinician | UWCH | WBF | | | | | no | 15 | 24 | 192 |
| **Sola et al. (2014 )** | Parent & Clinician | PPPM | FLACC | | | | | no | 27 | 12 | 60 |
| **Solodiuk (2013 )** | Parent, Clinician & Researcher | INRS | FLACC | PPP | PIPP | UWCH | NCCPC | no | 50 | 0 | 0 |
| **Solodiuk et al. (2010 )** | Parent, Clinician & Researcher | INRS | NCCPC | | | | | no | 50 | 72 | 216 |
| **Soltesz et al. (2007 )** | Researcher | OWN | | | | | | no | 64 | 24 | 72 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Somaini et al. (2016 )** | Researcher | FLACC | CHIPPS | CHEOPS | no | 512 | 12 | 72 |
| **Spanos et al. (2008 )** | Parent & Researcher | VAS | | | yes | 70 | 96 | 180 |
| **Spektor et al. (2016 )** | Parent | CHIPPS | | | no | 100 | 36 | 144 |
| **Splinter et al. (1995 )** | Clinician | CHEOPS | | | no | 156 | 18 | 156 |
| **Splinter et al. (1995 )** | Parent & Researcher | CHEOPS | VAS | | no | 202 | 12 | 156 |
| **Splinter et al. (1997 )** | Parent & Clinician | VAS | CHEOPS | | no | 164 | 24 | 72 |
| **Steib et al. (2005 )** | Clinician | CHEOPS | | | no | 38 | 30 | 72 |
| **St-Laurent-Gagnon et al. (1999)** | Parent | FAS | HPCT | MSPCT | yes | 104 | 48 | 72 |
| **Stoddard et al. (2006 )** | Clinician | WBF | | | no | | 12 | 48 |
| **Strout et al. (2011 )** | Clinician | PEPPS | | | no | 118 | 12 | 38 |
| **Strout et al. (2011)** | Clinician | MPEPPS | | | no | 118 | 12 | 36 |
| **Study et al. (2015 )** | Researcher | TDI | | | no | 943 | 18 | 36 |
| **Subhashini et al. (2008 )** | Parent & Clinician | WBF | CAS | | yes | 181 | 72 | 144 |
| **Suraseranivongse et al. (2003 )** | Clinician | CHEOPS | | | no | 103 | 12 | 144 |
| **Suresh et al. (2002 )** | Researcher | OPS | | | no | 40 | 24 | 216 |
| **Sutters et al. (1999 )** | Clinician | CHEOPS | | | yes | 87 | 36 | 144 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Sutters et al. (2012 )** | Parent | FLACC | | | no | 47 | 36 | 60 |
| **Sylvester et al. (2011 )** | Clinician | CHEOPS | | | no | 87 | 33 | 141 |
| **Taddio et al. (2009)** | Researcher & Clinician | VAS | MBPS | | no | 120 | 12 | 12 |
| **Taddio et al. (2017 )** | Parent, Clinician & Researcher | MBPS | NRS11 | | no | 296 | 15 | 15 |
| **Taheri et al. (2011 )** | Clinician | OPS | | | no | 60 | 36 | 144 |
| **Talu et al. (2008)** | Not named | OPS | | | no | 60 | 12 | 144 |
| **Tan et al. (1992)** | Clinician | VAS | | | yes | 73 | 60 | 180 |
| **Tarbell et al. (1992)** | Parent, Clinician & Researcher | TPPS | VAS | NRS | no | 74 | 12 | 64 |
| **Tay et al. (2012 )** | Parent | NRS10 | | | no | 41 | 12 | 180 |
| **Tazeroualti et al. (2007 )** | Researcher | OPS | | | no | 68 | 12 | 72 |
| **Tekelioglu et al. (2013 )** | Clinician | WBF | | | no | 60 | 48 | 120 |
| **Teo et al. (2011 )** | Clinician | FLACC | | | no | 52 | 24 | 190.8 |
| **Thompson et al. (2012 )** | Parent | PPPM | | | no | 202 | 48 | 216 |
| **Toker et al. (2016 )** | Clinician | CHIPPS | | | no | 75 | 24 | 84 |

|  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|
| **Trifa et al. (2009)** | Not named | CHEOPS |  |  | no | 72 | 36 | 108 |
| **Trifa et al. (2012 )** | Not named | CHEOPS |  |  | no | 60 | 12 | 72 |
| **Tripi et al. (2005 )** | Not named | WBF | FLACC |  | no | 35 | 12 | 120 |
| **Tsao et al. (2015 )** | Parent | WBF |  |  | yes | 59 | 36 | 144 |
| **Tsuchiya et al. (2004 )** | Parent | WBF |  |  | no | 30 | 12 | 96 |
| **Tuomilehto et al. (2000 )** | Clinician | MAUN |  |  | no | 100 | 12 | 108 |
| **Tuomilehto et al. (2002 )** | Clinician | MAUN |  |  | no | 120 | 12 | 108 |
| **Turan et al. (2003 )** | Clinician | TPPS |  |  | no | 44 | 12 | 60 |
| **Tyler et al. (1993 )** | Not named | CHEOPS |  | CRS5 | yes | 26 | 36 | 144 |
| **Ugur et al. (2013 )** | Not named | CHEOPS |  |  | no | 75 | 36 | 120 |
| **Ullan et al. (2014 )** | Not named | FLACC |  |  | no | 95 | 12 | 84 |
| **Umuroglu et al. (2004 )** | Researcher | CHEOPS |  |  | yes | 60 | 60 | 144 |
| **Usichenko et al. (2016 )** | Parent & Clinician |  | NRS11 |  | yes | 72 | 48 | 216 |
| **Uysal et al. (2011 )** | Clinician | OPS |  |  | no | 64 | 72 | 192 |
| **van der Putten et al. (2011 )** | Parent & Researcher | VAS | PBC |  | no | 16 | 36 | 84 |
| **van Dijk et al. (2001 )** | Clinician | COMFORT | VAS |  | no | 40 | 12 | 36 |
| **van Dijk et al. (2002 )** | Clinician | COMFORT | VAS |  | no | 35 | 12 | 36 |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| **Varghese et al. (2010 )** | Researcher | CRS4 |  | no | 84 | 60 | 180 |
| **Versloot et al. (2004 )** | Parent, Clinician & Researcher | CRS4 |  | no | 50 | 48 | 96 |
| **Vessey, et al. (1994)** | Researcher | CHEOPS |  | yes | 100 | 42 | 155 |
| **Vetter et al. (1996 )** | Parent & Clinician | NRS101 |  | yes | 30 | 96 | 192 |
| **Viitanen et al. (1999 )** | Clinician | OPS |  | no | 60 | 12 | 36 |
| **Viitanen et al. (1999 )** | Clinician | OPS |  | no | 52 | 12 | 36 |
| **Viitanen et al. (2000 )** | Clinician | OPS |  | no | 80 | 12 | 36 |
| **Viitanen et al. (2001 )** | Clinician | OPS |  | no | 80 | 12 | 36 |
| **Viitanen et al. (2003 )** | Clinician | OPS |  | no | 160 | 12 | 72 |
| **Voepel-Lewis et al. (2002)** | Parent, Clinician & Researcher | FLACC | VAS | yes | 79 | 48 | 216 |
| **von Baeyer et al. (2011 )** | Parent | PPPM |  | yes | 264 | 84 | 144 |
| **von et al. (2011 )** | Parent | FPSR |  | yes | 108 | 36 | 84 |
| **Walther-Larsen et al. (2016 )** | Parent | PPPM | NRS11 | no | 149 | 12 | 204 |
| **Watcha et al. (1992 )** | Researcher | OPS |  | yes | 95 | 60 | 180 |

|  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|
| **Wathen et al. (2007 )** | Researcher | FLACC | CHEOPS | yes | 120 | 15 | 216 |
| **Werk et al. (2008 )** | Clinician | OPS | FPSR | yes | 62 | 60 | 204 |
| **Wheeler et al. (2005 )** | Researcher | OPS |  | no | 30 | 24 | 103.2 |
| **Whitehead-Pleaux et al. (2006 )** | Clinician | NAPI |  | yes | 14 | 72 | 192 |
| **Wolf et al. (2002 )** | Clinician | CRS4 |  | yes | 125 | 48 | 144 |
| **Wong et al. (2015 )** | Parent |  | PPPM | yes | 33 | 48 | 72 |
| **Xiang et al. (2014 )** | Clinician | FLACC |  | no | 50 | 12 | 36 |
| **Yao et al. (2017)** | Researcher | FLACC |  | no | 28 | 12 | 72 |
| **Yeh et al. (2005 )** | Not named | FLACC |  | yes | 149 | 36 | 83.04 |
| **Yenigun et al. (2015 )** | Researcher | CHEOPS |  | no | 120 | 60 | 180 |
| **Yenigun et al. (2018)** | Researcher | CHEOPS |  | no | 63 | 24 | 168 |
| **Yildiz et al. (2010 )** | Clinician | CHIPPS |  | no | 63 | 12 | 84 |
| **Yilmaz et al. (2014 )** | Clinician | CHEOPS |  | no | 537 | 16 | 19 |
| **Yinger et al. (Winter, )** | Parent | UPAT |  | no | 58 | 48 | 72 |
| **Young, et al. (1988)** | Researcher | OWN |  | no | 80 | 48 | 83 |
| **Yu et al. (2015 )** | Clinician | CRS4 |  | no | 100 | 36 | 144 |
| **Zanchi et al. (2017)** | Parent & Clinician | NCCPC |  | no | 40 | 36 | 216 |
| **Zavras et al. (2014)** | Clinician | FLACC |  | no | 106 | 24 | 144 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Zempsky et al. (2003 )** | Parent, Clinician & Researcher | VAS | yes | 42 | 84 | 216 |
| **Zempsky et al. (2004 )** | Parent | WBF | yes | 86 | 60 | 132 |
| **Zempsky et al. (2008 )** | Parent | VAS | yes | 579 | 36 | 216 |
| **Zempsky et al. (2008 )** | Parent | VAS | yes | 60 | 36 | 84 |
| **Zhang et al. (2014 )** | Clinician | CHEOPS | yes | 61 | 48 | 192 |
| **Zhao et al. (2012 )** | Clinician | CAM4 | no | 192 | 36 | 120 |
| **Zhuang et al. (2011 )** | Clinician | CHEOPS | no | 60 | 24 | 156 |
| **Zisk et al. (2007 )** | Parent | PPPM | yes | 32 | 60 | 120 |